THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 59, No. 3, March 1980
Printed in U.S.A.

On Prediction of Moving-Average Processes

By L. A. SHEPP, D. SLEPIAN, and A. D. WYNER

(Manuscript received September 26, 1979)

Let $\{X_n\}$ be a discrete-time stationary moving-average process having the representation $X_n = \sum_1^M A_j Y_{n-j}$ where the real-valued process $\{Y_n\}$ has a well-defined entropy and spectrum. Let ϵ_k^{*2} denote the smallest mean-squared error of any estimate of X_n based on observations of $X_{n-1}, X_{n-2}, \cdots, X_{n-k}$, and let $\epsilon_{k\text{lin}}^{*2}$ be the corresponding least mean-squared error when the estimator is linear in the k observations. We establish an inequality of the form $\epsilon_\infty^{*2} > G(Y)\epsilon_{\infty\text{lin}}^{*2}$ where $G(Y) \leq 1$ depends only on the entropy and spectrum of $\{Y_n\}$. We also obtain explicit formulas for ϵ_k^{*2} and $\epsilon_{k\text{lin}}^{*2}$ and compare these quantities graphically when $M = 2$ and the $\{Y_n\}$ are i.i.d. variates with one of several different distributions. The best estimators are quite complicated but are frequently considerably better than the best linear ones. This extends a result of M. Kanter.

I. INTRODUCTION

This paper is concerned with the problem of estimating the current value, X_n, of a discrete-time stationary stochastic process given the k previous values, $X_{n-1}, X_{n-2}, \cdots, X_{n-k}$. Denote such an estimator, or predictor, by

$$\hat{X}_n = f_k(X_{n-1}, X_{n-2}, \cdots, X_{n-k}). \tag{1}$$

We adopt the mean-squared error

$$\epsilon_k^2 \equiv E[X_n - \hat{X}_n]^2 \tag{2}$$

as a figure of merit for the estimator f_k. Throughout the paper, we assume $EX_n = 0$, $n = 0, \pm 1, \pm 2, \cdots$.

It is well known, and easy to show, that no estimator has smaller mean-squared error than

$$f_k^*(X_{n-1}, \cdots, X_{n-k}) \equiv E(X_n \mid X_{n-1}, \cdots, X_{n-k}), \tag{3}$$

the conditional expectation of X_n given $X_{n-1}, \cdots, X_{n-k}$. We denote the

mean-squared error of this best estimator by

$$\epsilon_k^{*2} \equiv E[X_n - f_k^*]^2 = E[X_n - E(X_n \mid X_{n-1}, \cdots, X_{n-k})]^2. \tag{4}$$

While the best estimator is simply described by (3), in practice it is frequently impossible to calculate it explicitly for processes of interest.

The simpler class of linear estimators

$$f_{k\text{lin}}(X_{n-1}, \cdots, X_{n-k}) = \sum_1^k c_{kj} X_{n-j} \tag{5}$$

has been much studied in the past.[1] It is well known how to choose the c's to obtain the smallest mean-squared error within this class of predictors, and this least mean-squared error is given by the simple formula

$$\epsilon_{k\text{lin}}^{*2} = D_{k+1}/D_k, \tag{6}$$

where D_l is the $l \times l$ determinant whose entry in the ith row and jth column is

$$\rho_{i-j} \equiv E[X_i - EX_i][X_j - EX_j] \tag{7}$$

$i, j = 1, 2, \cdots, l$. The optimizing c's are also given by determinants involving the ρ_{i-j} so that all quantities of interest for the optimal linear predictor are specified by the second-order statistics of the process and are generally easy to compute explicitly. For this reason, if for no other, the optimal linear estimator bas been much studied and used in practice.

How does ϵ_k^{*2} compare with $\epsilon_{k\text{lin}}^{*2}$? How much does nonlinear estimation buy? The answer, of course, depends on the process $\{X_n\}$. On one hand, for Gaussian processes $\epsilon_k^{*2} = \epsilon_{k\text{lin}}^{*2}$, so nothing is gained; on the other, one can construct processes for which $\epsilon_{k\text{lin}}^{*2}/\epsilon_k^{*2}$ is arbitrarily large.

In this paper we study predictors for some moving-average processes of the form

$$X_n = \sum_{j=0}^M A_j Y_{n-j}. \tag{8}$$

These processes are often used as models in applications. When the Y's are identically distributed independent random variables, the X process is sometimes called "filtered white noise." We establish for (8) a quite general bound of the form

$$\epsilon_\infty^{*2} \geq G(Y)\epsilon_{k\text{lin}}^{*2}, \tag{9}$$

where the constant $G(Y) \leq 1$ depends only on the spectrum and entropy of the Y process and is independent of the A's of eq. (8). When the Y's are independent identically distributed (i.i.d.) random vari-

ables, $G(Y)$ is particularly simple to compute. Our bound generalizes a similar one found by Kanter.[2]

We are able to find f_k^* and ϵ_k^{*2} explicitly for several special cases of (8). We treat in complete detail the case

$$X_n = A_0 Y_n + A_1 Y_{n-1}, \tag{10}$$

where the Y's are either i.i.d. with a uniform distribution on an interval or are i.i.d. with a one-sided exponential distribution. Curves are presented in these cases that compare ϵ_k^{*2} with $\epsilon_{k\text{lin}}^{*2}$ for various values of the parameters involved. It is interesting to note here that even for $k = 1$, $\epsilon_k^{*2} < \epsilon_{k\text{lin}}^{*2}$ for a wide range of parameter values. The explicit results are compared with the bounds already mentioned.

Another special case of (10) is worked out in detail. Here $A_0 = 1$, $A_1 = \pm 1$ and the Y_n are i.i.d. with the discrete distribution $\Pr[Y_n = 1] = p = 1 - \Pr[Y_n = -1]$.

We are also able to exhibit the best predictor when

$$X_n = \sum_{j=0}^{M} \alpha^j Y_{n-j} \tag{11}$$

and the Y's are i.i.d. uniform or have a one-sided exponential distribution. The best predictors are surprisingly complicated here. We obtained an expression for the least error, ϵ_k^{*2}, but it is not included as it is apparently useless even for numerical calculation.

Our results are presented in detail in the next section. Derivations, proofs, and further discussion of general theory are relegated to the succeeding sections.

II. RESULTS

Let

$$X_n = Y_n - aY_{n-1}, \quad EY_n = 0, \quad EY_n^2 = 1, \tag{12}$$

$n = 0, \pm 1, \pm 2, \cdots$ where the Y's are independent and identically distributed random variables. Then

$$\rho_j \equiv EX_n X_{n-j} = \begin{cases} 1 + a^2, & j = 0 \\ -a, & |j| = 1 \\ 0, & |j| > 1 \end{cases} \tag{13}$$

and the determinant D_k of (6) has value $(1 - a^{2(k+1)})/(1 - a^2)$ so that the figure of merit for the best linear predictor is

$$\epsilon_{k\text{lin}}^{*2} = \frac{1 - a^{2k+4}}{1 - a^{2k+2}} \underset{k\to\infty}{\to} \begin{cases} 1, & |a| \le 1 \\ a^2, & |a| > 1, \end{cases} \tag{14}$$

a result that does not depend on the distribution of Y_n. From (14) we see that the behavior of the best linear predictor as $k \to \infty$ depends

markedly on whether or not $|a| < 1$. Since, from (12), Y_n has unit variance and is independent of $Y_{n-1}, X_{n-1}, X_{n-2}, \cdots$, it follows that $\epsilon_k^{*2} \geq \text{var } Y_n = 1$ for $k = 1$. Thus, from (14) and the fact that always $\epsilon_k^{*2} \leq \epsilon_{k\text{lin}}^{*2}$, we have

$$\epsilon_{\infty\text{lin}}^{*2} = 1 = \epsilon_\infty^{*2}, \qquad |a| \leq 1. \tag{15}$$

When Y_n in (12) has the uniform distribution with density

$$p_Y(y) = \begin{cases} \dfrac{1}{2\gamma}, & |y| \leq \gamma \\ 0, & |y| > \gamma \end{cases}$$
$$\gamma \equiv \sqrt{3}, \tag{16}$$

we find that

$$E(X_n | X_{n-1}, \cdots, X_{n-k}) \equiv f_k^*(X_{n-1}, X_{n-2}, \cdots, X_{n-k}; a)$$
$$= \frac{a}{2}[\max(-\gamma, U_1, U_2, \cdots, U_k) + \min(\gamma, V_1, V_2, \cdots, V_k)], \qquad a \geq 0, \tag{17}$$

where

$$U_i = \sum_{j=1}^{i} a^{j-1} X_{n-j} - a^i \gamma,$$
$$V_i = \sum_{j=1}^{i} a^{j-1} X_{n-j} + a^i \gamma,$$
$$i = 1, 2, \cdots, k. \tag{18}$$

When $a \equiv -b < 0$,

$$E(X_n | X_{n-1}, \cdots, X_{n-k}) = f_k^*(-X_{n-1}, X_{n-2}, -X_{n-3}, \cdots, (-1)^k X_{n-k}; b). \tag{19}$$

For all values of a, the mean-squared error of this best predictor is

$$\epsilon_k^{*2} = 1 + 6a^2 \int_0^1 du\, u(1-u) \prod_{i=1}^{k} \left(1 - \frac{u}{|a|^i}\right)^+ \tag{20}$$

where $(x)^+ = x$ if $x \geq 0$ and is zero otherwise. This can be written in the alternative form

$$\epsilon_k^{*2} = \begin{cases} 1 + 6a^{2k+2} P_k(|a|), & |a| \leq 1 \\ 1 + 6a^2 P_k\left(\dfrac{1}{|a|}\right), & |a| \geq 1 \end{cases} \tag{21}$$

where the polynomial $P_k(x)$ of degree $\binom{k}{2}$ in x is given explicitly by

$$P_k(x) = \frac{1}{6} + \sum_{l=1}^{k} \frac{(-1)^l x^{l(l+1)/2}}{(l+2)(l+3)} \prod_{j=1}^{l} \frac{1 - x^{k+1-j}}{1 - x^j}. \tag{22}$$

Another form, more suitable for computation, is

$$P_k(x) = \sum_{0}^{k} \frac{b_j}{(j+2)(j+3)}, \quad b_0 = 1, \quad b_{j+1} = \frac{x^{k+1} - x^{j+1}}{1 - x^{j+1}} b_j. \tag{23}$$

Figure 1 shows $\epsilon_1^{*2}/\epsilon_{\infty\text{lin}}^{*2}$ vs a for the case at hand. When $|a| > 2$, the best estimate of X_n based on just X_{n-1} has slightly smaller mean-squared error than the best linear estimator based on the infinite past. As is seen, as k increases, ϵ_k^{*2} is smaller than $\epsilon_{\infty\text{lin}}^{*2}$ for a large range of a values.

Figure 2 compares the best estimator based on k past samples of X with the best linear estimator based on the same samples. For $|a| < 1$ and large k, the linear estimator does nearly as well as the best; but for $a \approx 1.5$ the nonlinear estimator (17) is significantly better. The curves shown approach the limit indicated as $k \to \infty$. For $a > 1.2$ to the scale shown on Fig. 2, it coincides with the curve labeled $k = 20$. The bound (9) gives $\epsilon_\infty^{*2}/\epsilon_{\infty\text{lin}}^{*2} \geqq 6/\pi e = 0.703$, a value about 14 percent less than the minimum shown on Fig. 2.

As noted in (13), the process (12) always has variance $EX_n^2 = 1 + a^2$. Figure 3 shows the mean-squared error of the best predictor for a process of form (12) scaled to have unit variance. Again the Y's are i.i.d. with a uniform distribution. The limiting curve for $k \to \infty$ has the value $1/(1 + a^2)$ for $|a| \leq 1$ and, to the scale shown, coincides with the curve labeled $k = 40$ for $a > 1$.

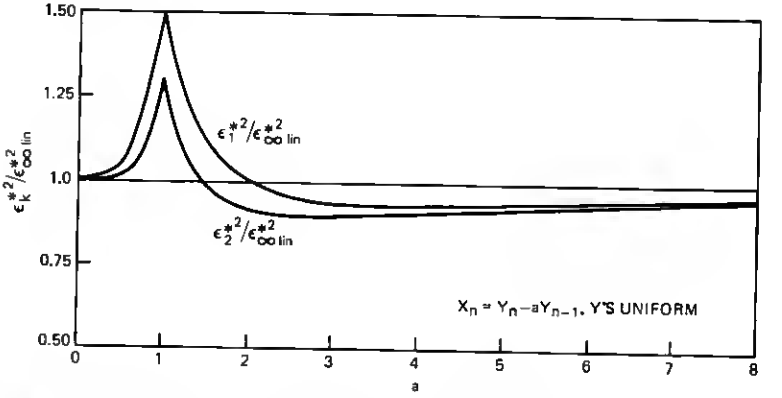


Fig. 1—Comparison of ϵ_1^{*2} and ϵ_2^{*2} with $\epsilon_{\infty\text{lin}}^{*2}$ for $X_n = Y_n - aY_{n-1}$ with the Y's i.i.d. uniform variates.

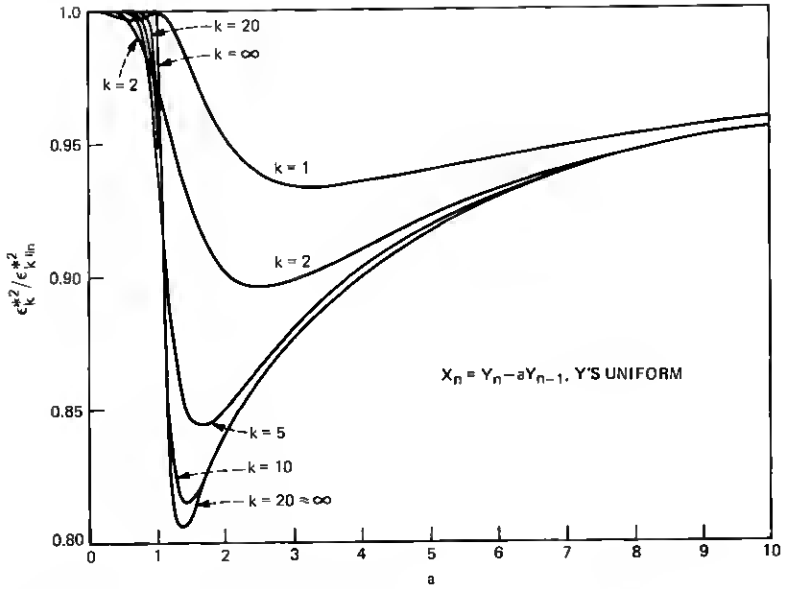


Fig. 2—Comparison of ϵ_k^{*2} with $\epsilon_{k\text{lin}}^{*2}$ for $X_n = Y_n - aY_{n-1}$ with the Y's i.i.d. uniform variates.

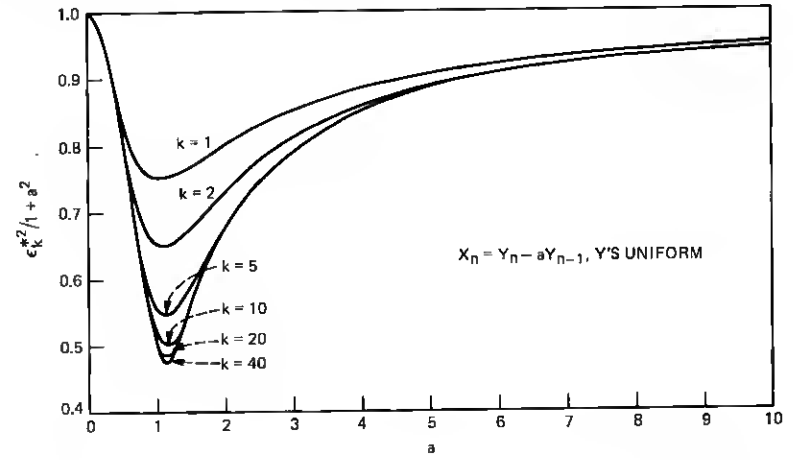
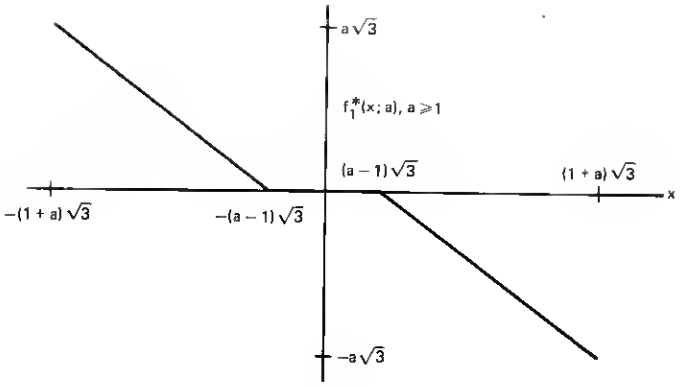


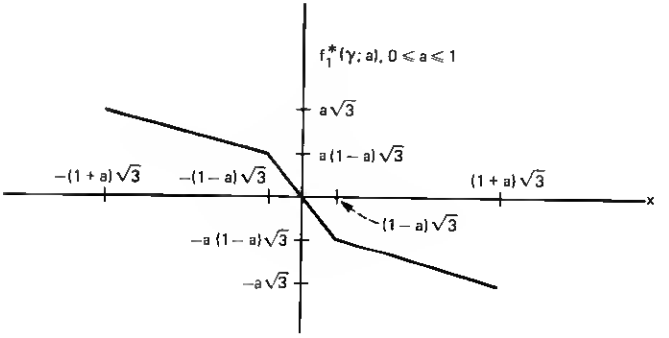
Fig. 3—$\epsilon_k^{*2}/(1 + a^2)$ vs a for $X_n = Y_n - aY_{n-1}$ with the Y's i.i.d. uniform variates.

It is instructive to examine (17) more closely to understand the nonlinear nature of the best estimator. Figure 4a shows $f_1^*(x; a)$ as a function of x for $a > 1$; Fig. 4b shows this quantity when $0 \leq a \leq 1$. Consider the case where $a > 1$. Now

$$X_{n-1} = Y_{n-1} - aY_{n-2}$$



(a)



(b)

Fig. 4—The best estimator $f_1^*(x; a)$ for the process $X_n = Y_n - aY_{n-1}$ with the Y's i.i.d. uniform variates.

and each $|Y| \leq \gamma$. If an observation of X_{n-1} has the value $(1 + a)\gamma$, then we must have $Y_{n-1} = \gamma$ and $Y_{n-2} = -\gamma$. Then $X_n = Y_n - aY_{n-1} = Y_n - a\gamma$. Since Y_n has a symmetric distribution, the best estimate of X_n is now its mean, $-a\gamma$. If now, instead of observing X_{n-1} at the extreme value $(1 + a)\gamma$, we observe a value x near the extreme, say where $(a - 1)\gamma < x < (1 + a)\gamma$, we still obtain some information about Y_{n-1}. In fact, one easily calculates for x in this range that

$$p_{Y_{n-1}|X_{n-1}}(y|x) = \begin{cases} \dfrac{1}{(1+a)\gamma - x}, & x - a\gamma < y < \gamma \\ 0, & \text{otherwise} \end{cases}$$

and that $E[Y_{n-1}|X_{n-1} = x] = \frac{1}{2}[x + (1 - a)\gamma]$. Then, from (12),

$$E(X_n \mid X_{n-1} = x) = E(Y_n \mid X_{n-1} = x)$$

$$- aE(Y_{n-1} \mid X_{n-1} = x) = -\frac{a}{2}[x + (1 - a)\gamma].$$

When $|x| < (a - 1)\gamma$, $p_{Y_{n-1}|X_n}(y|x) = p_{Y_{n-1}}(y)$ and knowledge of X_{n-1} no longer gives information about Y_{n-1}. The best estimate of X_n is then its mean which is zero. The case $|a| < 1$ can be discussed in a similar manner. When k, the number of past X values observed, is larger than 1, this sort of analysis becomes difficult, however, and the intuitive understanding of (17) becomes obscure.

Figures 5, 6, 7, and 8 give results similar to those of Figs. 2 and 3, but for the case in which the Y's of (12) are i.i.d. with density,

$$p_Y(y) = \begin{cases} e^{-(y+1)}, & y \geq -1 \\ 0, & y < -1. \end{cases} \tag{24}$$

For $a \geq 0$, we find

$$f_k^* = -a[\delta + \max[-1, W_1, W_2, \cdots, W_k]] \tag{25}$$

where

$$W_i = \sum_{j=1}^{i} a^{j-1} X_{n-j} - a^i, \quad i = 1, 2, \cdots, k,$$

$$\delta = \frac{a^k(1 - a)}{1 - a^{k+1}}, \tag{26}$$

and the figure of merit for this best estimator is

$$\epsilon_k^{*2} = 1 + a^2\delta^2. \tag{27}$$

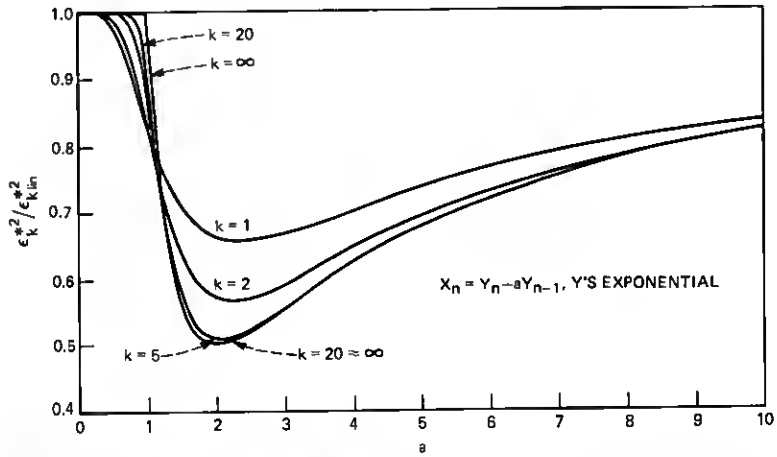


Fig. 5—Comparison of ϵ_k^{*2} with $\epsilon_{k\,lin}^{*2}$ for $a > 0$ for $X_n = Y_n - aY_{n-1}$ with the Y's i.i.d. one-sided exponential variates.

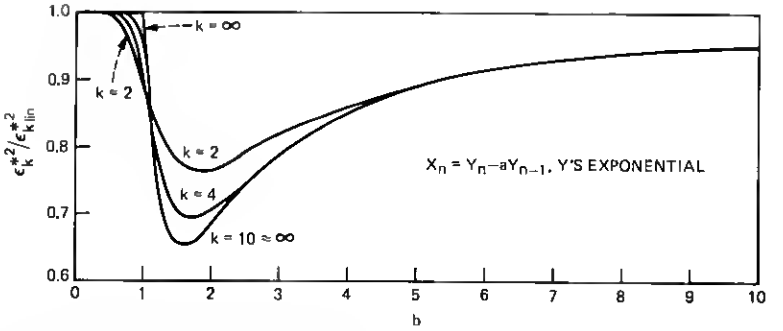


Fig. 6—Comparison of ϵ_k^{*2} with $\epsilon_{k\mathrm{lin}}^{*2}$ for $a = -b < 0$ for $X_n = Y_n - aY_{n-1}$ with the Y's i.i.d. one-sided exponential variates.

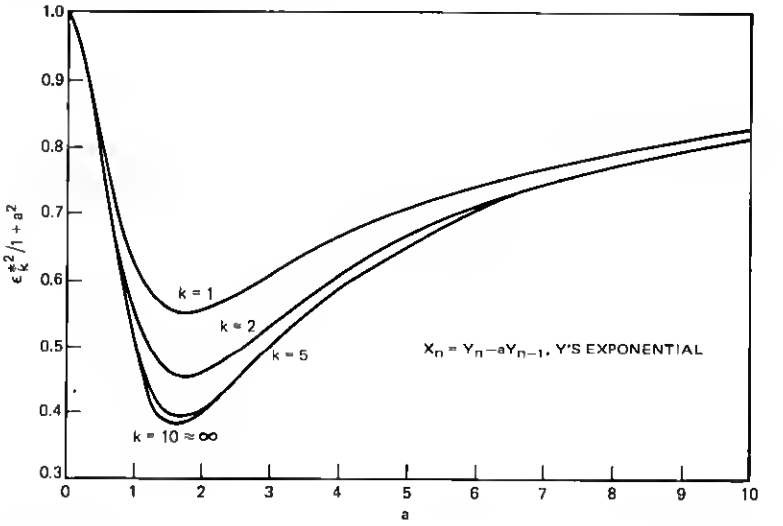


Fig. 7—$\epsilon_k^{*2}/(1 + a^2)$ vs $a > 0$ for $X_n = Y_n - aY_{n-1}$ with the Y's i.i.d. one-sided exponential variates.

For $a = -b < 0$, the estimator is given by the more complicated expression

$$f_k^* = -a\left[\delta + \frac{Ae^{-\delta A} - Be^{-\delta B}}{e^{-\delta A} - e^{-\delta B}}\right],$$

$$A \equiv \max[-1, W_2, W_4, \cdots, W_{k_e}],$$

$$B \equiv \min[W_1, W_3, \cdots, W_{k_0}], \tag{28}$$

$$k_e = \begin{cases} k, & k \text{ even} \\ k-1, & k \text{ odd} \end{cases}, \qquad k_0 = \begin{cases} k, & k \text{ odd} \\ k-1, & k \text{ even.} \end{cases}$$

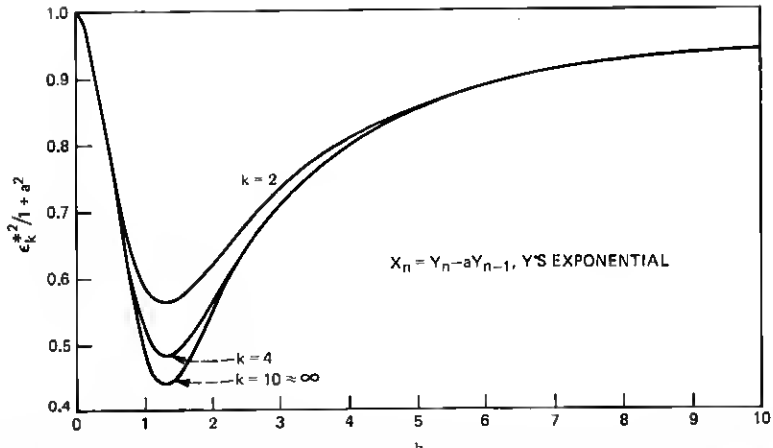


Fig. 8—$\epsilon_k^{*2}/(1 + a^2)$ vs $a = -b < 0$ for $X_n = Y_n - aY_{n-1}$ with the Y's i.i.d. one-sided exponential variates.

Its mean-squared error is given by

$$\epsilon_k^{*2}(b) = 1 + \left(\frac{b^{k+1}(1+b)}{1-(-b)^{k+1}}\right)^2 \left[1 - 2\lambda(1+\lambda)\sum_{n=1}^{\infty}\frac{1}{(1+n\lambda)^3}\right],$$

$$\lambda = \begin{cases} (b-1)(b^{k+1}+1)/(b(b^k-1)), & k = 2, 4, 6, \cdots \\ b - 1,\ b > 1 \quad , & k = 1, 3, 5, \cdots \\ \dfrac{1}{b} - 1,\ b < 1 \end{cases}$$

$$\epsilon_k^{*2}(1) = 1 + \frac{2}{(k+1)^2}, \quad k = 1, 3, \cdots. \tag{29}$$

It is seen from the curves that, for $a > 0, \epsilon_k^{*2}(a) \leq \epsilon_k^{*2}(-a)$, at least for the graphs drawn. This raises the question of comparing $\epsilon_k^{*2}(a)$ and $\epsilon_k^{*2}(-a)$ in general. It is easy to see (Appendix A) that these are equal if Y has a symmetric distribution. We show in Appendix B that $\epsilon_k^{*2}(a, Y) \leq \epsilon_k^{*2}(-a, Y)$ for every Y if $k = 1$ and $a = 1$. However, the inequality is false in general for $k = 1$ if $a \neq 1$. The case $k = 1$, $a = 1$ is thus special and the inequality is shown there to be the same as the fact that, if Y_0 and Y_1 are i.i.d. variates, then the average conditional variance of Y_0 given $Y_0 - Y_1$ is smaller than the average conditional variance of Y_0 given $Y_0 + Y_1$.

The bound (9) for the present case yields $\epsilon_\infty^{*2}/\epsilon_{\infty\text{lin}}^{*2} \geq e/2\pi = 0.4326$, a value 14 percent less than the minimum shown on Fig. 5.

Consider now the case in which the Y's of (12) have the discrete distribution

$$\Pr[Y_n = \lambda) = p, \qquad \Pr[Y_n = \mu] = q = 1 - p, \qquad \lambda < \mu. \tag{30}$$

To satisfy $EY_n = 0$ and $EY_n^2 = 1$, we must have

$$\lambda = -\sqrt{q/p}, \qquad \mu = \sqrt{p/q}. \tag{31}$$

We assume $a \neq 0$. Each X_n can then take only four possible values: $\mu - a\mu$, $\mu - a\lambda$, $\lambda - a\mu$, $\lambda - a\lambda$. If $a \neq \pm 1$, these four values are distinct. In this case, observation of X_{n-1} allows one to deduce the value of Y_{n-1}. The best estimate of X_n is then a times this value of Y_{n-1}, and this estimator has figure of merit $\epsilon_1^{*2} = 1$, the least value possible. The best linear estimator still has variance given by (14) and, for $|a| > 1$, this can be arbitrarily large. If $a = \pm 1$, however, the values of X_{n-1} no longer determine Y_{n-1}, and the best estimator is more complicated. It is,† for $a = 1$,

$$f_k^* = \begin{cases} -(Z_k + \lambda), & \text{at least one } Z = \mu - \lambda \\ -(Z_k + \mu), & \text{at least one } Z = \lambda - \mu \\ -\dfrac{\lambda p^{k+1} + \mu q^{k+1}}{p^{k+1} + q^{k+1}}, & \text{all } Z\text{'s} = 0 \end{cases} \tag{32}$$

and, for $a = -1$,

$$f_k^* = Z_k + (-1)^k \begin{cases} \dfrac{\lambda p^{s+1} q^t + q^{s+1} p t}{p^{s+1} q^t + q^{s+1} p^t}, & \text{all } Z\text{'s either } 0 \text{ or } \lambda + \mu \\ \lambda, & \text{at least one } Z \text{ either } 2\lambda \text{ or } \mu - \lambda \\ \mu, & \text{at least one } Z \text{ either } 2\mu \text{ or } \lambda - \mu \end{cases} \tag{33}$$

where

$$Z_i \equiv \sum_{j=1}^{i} a^{j-1} X_{n-k+(i-j)}, \quad i = 1, 2, \cdots, k,$$

$$s = \text{number of positive even integers} \leq k,$$

$$t = \text{number of positive odd integers} \leq k. \tag{34}$$

The figure of merit for these estimators is

$$\epsilon_k^{*2} = 1 + \frac{p^k q^k}{p^{k+1} + q^{k+1}}, \qquad a = 1 \tag{35}$$

$$\epsilon_k^{*2} = 1 + \begin{cases} (pq)^{k/2}, & k \text{ even} \\ \tfrac{1}{2}(pq)^{(k-1)/2}, & k \text{ odd} \end{cases} \qquad a = -1. \tag{36}$$

† It turns out that the three alternatives in (32) and (33) are mutually exclusive and exhaustive, respectively.

The best linear estimator in this case for $a = \pm 1$ has mean-squared error

$$\epsilon_{k\text{lin}}^{*2} = 1 + \frac{1}{k+1}. \tag{37}$$

Figure 9 shows the best estimator for this case.

The case (12) just treated is indeed as general as (10) in which the i.i.d. Y's have finite variance σ^2 and mean m. If $f_k^*(X_{n-1}, \cdots, X_{n-k}; a)$ and $\epsilon_k^{*2}(a)$ are for (12) the best estimator and its figure of merit, then the best estimator for X_n of (10) is

$$A_0\sigma f_k^*(\bar{X}_{n-1}, \cdots, \bar{X}_{n-k}; -A_1/A_0) + m(A_0 + A_1)$$

where $\bar{X}_j \equiv [X_j - (A_0 + A_1)m]/(A_0\sigma)$ for all j, and the figure of merit for this best estimator is $A_0^2\sigma^2\epsilon_k^{2*}(-A_1/A_0)$.

As for the inequality (9) mentioned in Section I, we show in Section VI that when X_n is given by (8)

$$\epsilon_\infty^{*2} \geq \epsilon_{\infty\text{lin}}^{*2}\left[\frac{1}{2\pi e}\frac{e^{2H(Y)}}{\tilde{S}_Y}\right],$$

where $H(Y)$ is the differential entropy of the process Y (defined in (156)), and $\tilde{S}_Y = \exp\{\int_0^1 \log S_Y(f)\, df\}$ is the geometric mean of the spectral density $S_Y(f)$, defined in (178). The heart of the result is Theorem 2, in Section 6, which relates the entropy $H(X)$ to the entropy $H(Y)$.

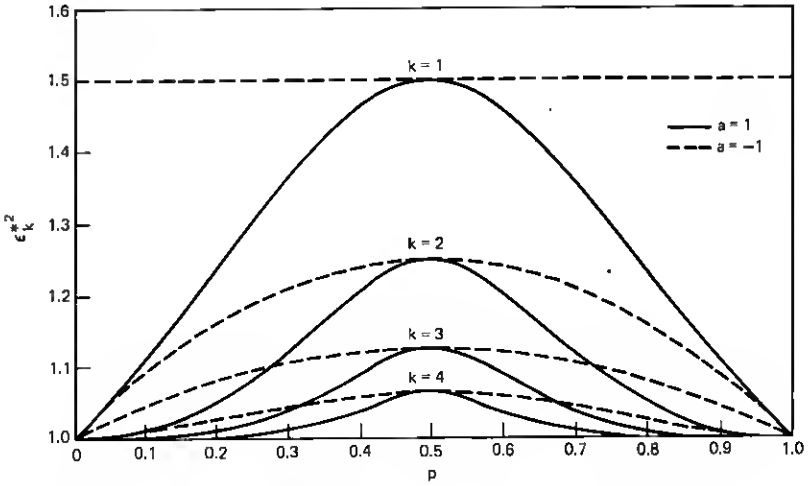


Fig. 9—ϵ_k^{*2} for the discrete valued process $X_n = Y_n - aY_{n-1}$ where the Y's are i.i.d. variates taking two values $\Pr[Y_n = -\sqrt{q/p}] = p = 1 - \Pr[Y_n = \sqrt{p/q}] = 1 - q$.

III. GENERAL THEORY

We consider the moving-average process

$$X_n = Y_n + a_1 Y_{n-1} + \cdots + a_M Y_{n-M} = \sum_j a_j Y_{n-j}$$

$$EY_n = 0, \quad EY_n^2 = 1, \quad a_0 \equiv 1, \quad a_j \equiv 0, \quad \begin{matrix} j > M \\ j < 0 \end{matrix}$$

$$n = 0, \pm 1, \pm 2, \cdots. \tag{38}$$

Let

$$R^{(l)} = \begin{vmatrix} 1 & a_1 & a_2 & \cdots & \cdot \\ 0 & 1 & a_1 & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{vmatrix}_{l \times l} \tag{39}$$

be the $l \times l$ upper-triangular matrix with element a_{j-i} in the ith row and jth column, $i, j = 1, 2, \cdots, l$. We adopt the notation

$$\mathbf{X}_i^j = (X_j, X_{j-1}, \cdots, X_i)^T$$

for the column vector whose components from top to bottom are X_j, $X_{j-1}, \cdots, X_i$. Then from (38) we can write

$$\mathbf{X}_{n-k}^n = R^{(k+1)} \mathbf{Y}_{n-k}^n + \mathbf{W}^{(n,k+1)} \tag{40}$$

where the components of the $(k+1)$-column vector $\mathbf{W}^{(n,k+1)}$ are, from top to bottom,

$$W_i^{(n,k+1)} = \sum_{l=1}^{\infty} a_{k+l+1-i} Y_{n-k-l},$$

$$i = 1, 2, \cdots, k+1. \tag{41}$$

Now denote the inverse of $R^{(l)}$ by

$$S^{(l)} = R^{(l)-1} = (S_{ij}^{(l)})_{l \times l}. \tag{42}$$

By multiplying (40) by $S^{(k+1)}$, we find

$$S^{(k+1)} \mathbf{X}_{n-k}^n = \mathbf{Y}_{n-k}^n + S^{(k+1)} \mathbf{W}^{(n,k+1)}. \tag{43}$$

The first component of (43) yields

$$\sum_{j=1}^{k+1} S_{1j}^{(k+1)} X_{n+1-j} = Y_n + \sum_{j=1}^{k+1} S_{1j}^{(k+1)} W_j^{(n,k+1)}$$

$$n = 0, \pm 1, \cdots, \quad k = 1, 2, \cdots.$$

Using (41), we obtain the useful result

$$\sum_{j=1}^{k+1} S_{1j}^{(k+1)} S_{n+1-j} = Y_n + \sum_{j=1}^{M} A_j^{(k+1)} Y_{n-k-j}$$

$$n = 0, \pm 1, \cdots, \quad k = 1, 2, \cdots \tag{44}$$

where

$$A_l^{(k+1)} \equiv \sum_{j=1}^{k+1} S_{1j}^{(k+1)} a_{k+1+l-j}$$

$$l = 1, 2, \cdots, M. \tag{45}$$

From (44) we see that

$$X_n = Y_n - \sum_{j=2}^{k+1} S_{1j}^{(k+1)} X_{n+1-j} + \sum_{j=1}^{M} A_j^{(k+1)} Y_{n-k-j}, \tag{46}$$

since clearly $S_{11}^{(k+1)} = 1$. From (46) it follows that

$$f_k^*(\mathbf{X}_{n-k}^{n-1} = E(X_n | \mathbf{X}_{n-k}^{n-1})$$

$$= -\sum_{j=2}^{k+1} S_{1j}^{(k+1)} X_{n+1-j} + \sum_{l=1}^{M} A_l^{(k+1)} E(Y_{n-k-l} | \mathbf{X}_{n-k}^{n-1}). \tag{47}$$

Before proceeding further with the calculation of the expectations on the right of (47), we comment on the special form of the matrix $S^{(l)}$ Consider the quantities d_l, $l = 1, 2, \cdots$ defined by

$$d_l \equiv (-1)^l \begin{vmatrix} a_1 & a_2 & a_3 & \cdot & \cdot & \cdot \\ 1 & a_1 & a_2 & \cdot & \cdot & \cdot \\ 0 & 1 & a_1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & a_1 & a_2 \\ 0 & 0 & 0 & \cdot & 1 & a_1 \end{vmatrix}_{l \times l} \tag{48}$$

where the determinant on the right has entries constant along diagonals as indicated. Expanding the determinant by the elements of the first row, we see that

$$d_l = -\sum_{j=1}^{l} a_j d_{l-j}, \quad d_0 \equiv 1,$$

$$l = 1, 2, \cdots. \tag{49}$$

We extend the definition of the d's by $d_j \equiv 0, j < 0$. It follows easily then by direct matrix multiplication that

$$S^{(l)} = \begin{vmatrix} 1 & d_1 & d_2 & \cdots & d_{l-1} \\ 0 & 1 & d_1 & \cdots & d_{l-2} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{vmatrix}_{l \times l} \tag{50}$$

is indeed the inverse of $R^{(l)}$ displayed in (39). Thus we have

$$S_{ij}^{(l)} = d_{j-i}$$

$$i, j = 1, 2, \cdots, l \tag{51}$$

and the quantities $A_l^{(k)}$ of (44) are

$$A_l^{(k)} = \sum_{j=1}^{k} a_{j+l-1} d_{k-j}$$

$$l = 1, 2, \cdots, M. \tag{52}$$

The asymptotic behavior of the $A_l^{(k)}$ as $k \to \infty$ is readily seen from this form. Since the d's satisfy the linear recurrence (49),

$$d_l = \sum_{1}^{M} D_i \alpha_i^l \tag{53}$$

where the quantities $\alpha_1, \alpha_2, \cdots, \alpha_M$ are the reciprocals of the roots (here assumed distinct) of the polynomial

$$Q(z) = \sum_{0}^{M} a_j z^j. \tag{54}$$

In terms of these roots, (52) becomes

$$A_l^{(k)} = \sum_{j=1}^{k} a_{j+l-1} \sum_{i=1}^{M} D_i \alpha_i^{k-j}.$$

Now let $k = M + p$, $p > 0$. Since $a_i = 0$ for $i > M$, we have

$$A_l^{(M+p)} = \sum_{j=1}^{M} a_{j+l-1} \sum_{i=1}^{M} D_i \alpha_i^{M+p-i}$$

$$= \sum_{i=1}^{M} D_i \alpha_i^p \sum_{j=1}^{M} a_{M+l-j} \alpha_i^{j-1}. \tag{55}$$

Since the inner sum here is independent of p, we see that

$$|\alpha_i| < 1, \quad i = 1, 2, \cdots, M \Rightarrow |A_l^{(k)}| \underset{k \to \infty}{\to} 0, \quad l = 1, 2, \cdots, M. \tag{56}$$

Now (46) shows X_n to be a linear combination of Y_n (which is independent of past X's), the expression

$$\hat{X}_k = -\sum_{j=2}^{k+1} S_{1j}^{(k+1)} X_{n+1-j} = -\sum_{j=1}^{k} d_j X_{n-j} \tag{57}$$

which is linear in past X's (we have used (51)), and the random variable $\sum_1^M A_j^{(k+1)} Y_{n-k-j} \equiv \tilde{X}$. We have $E\tilde{X} = 0$, $E\tilde{X}^2 = \sum_1^M [A_j^{(k+1)}]^2$, and this quantity approaches zero with increasing k if the $|\alpha_i|$ are all less than unity. Thus we see that in this case $E(X_n - \hat{X}_k)^2 \to 1$ as $k \to \infty$.

Theorem 1: If the roots of

$$\sum_{j=0}^{M} a_j z^j = 0$$

are all of magnitude greater than unity, a best estimator based on the infinite past is the linear form

$$f_\infty^*(X_{n-1}, X_{n-2}, \cdots) = -\sum_{j=1}^{\infty} d_j X_{n-j}$$

and

$$\epsilon_\infty^{*2} = 1 = \epsilon_{\infty \text{lin}}^{*2}.$$

Note that the linear estimator $\hat{X}_k$ of (57) is not the best linear estimator for any finite k. However, unlike the coefficients c_{kj} of the best linear estimator (5), the nonzero coefficients in (57) do not depend on k.

We return now to the calculation of f_k^* as given by (47). Rewrite (43) replacing n by $n-1$ and k by $k-1$ to obtain

$$S^{(k)}\mathbf{X}_{n-k}^{n-1} = \mathbf{Y}_{n-k}^{n-1} + S^{(k)}\mathbf{W}^{(n-1,k)}$$

or in component form

$$\sum_{j=1}^{k} S_{ij}^{(k)} X_{n-j} = Y_{n-i} + \sum_{j=1}^{k} S_{ij}^{(k)} \sum_{l=1} a_{k+l-j} Y_{n-k-l}$$

$$i = 1, 2, \cdots, k$$

or

$$Z_{k+1-i} \equiv \sum_{j=1}^{k} d_{j-i} X_{n-j} = Y_{n-i} + \sum_{j=1}^{M} A_j^{(k+1-i)} Y_{n-k-j},$$

$$i = 1, 2, \cdots, k, \tag{58}$$

the A's being given as before by (45). The triangular matrix $S^{(k)}$ connecting $Z_k, Z_{k+1}, \cdots, Z_1$ and the observed $X_{n-1}, X_{n-2}, \cdots, X_{n-k}$ via $\mathbf{Z}_1^k = S^{(k)}\mathbf{X}_{n-k}^{n-1}$ is nonsingular, so that in (47) we can now write

$$E(Y_{n-k-l} | \mathbf{X}_{n-k}^{n-1}) = E(Y_{n-k-l} | \mathbf{Z}_1^k). \tag{59}$$

Now

$$E(Y_{n-k-l} | \mathbf{Z}_1^k) = \int dy_1 \cdots \int dy_M \, y_l p_{\mathbf{Y}_{n-k-1}^{n-k-M} | \mathbf{Z}_1^k}(\mathbf{y}_1^M | \mathbf{Z}_1^k). \tag{60}$$

But, by Bayes' rule for conditional probabilities,

$$p_{\mathbf{Y}_{n-k-1}^{n-k-M} | \mathbf{Z}_1^k} = \frac{p_{\mathbf{Z}_1^k | \mathbf{Y}_{n-k-1}^{n-k-M}} p_{\mathbf{Y}_{n-k-1}^{n-k-M}}}{p_{\mathbf{Z}_1^k}}. \tag{61}$$

Denote the density of Y_j by

$$p_{Y_j}(y) = g(y). \tag{62}$$

Since the Y's are i.i.d.,

$$p_{\mathbf{Y}_{n-k-1}^{n-k-M}}(\mathbf{y}_1^M) = \prod_{i=1}^{M} g(y_i). \tag{63}$$

Furthermore, we see from (58) that given $\mathbf{Y}_{n-k-1}^{n-k-M} = \mathbf{y}_1^M$ the Z's are independent random variables, so that

$$p_{\mathbf{Z}_1^k|\mathbf{Y}_{n-k-1}^{n-k-M}}(Z_1^k|\mathbf{y}_1^M) = \prod_{l=1}^{k} g\left(Z_l - \sum_{j=1}^{M} A_j^{(l)} y_j\right). \tag{64}$$

Equations (59) to (64) combine to yield

$$E(Y_{n-k-l}|\mathbf{X}_{n-k}^{n-1}) = \frac{I_{l1}(\mathbf{Z}_1^k)}{I_2(\mathbf{Z}_1^k)}, \tag{65}$$

$$I_{l1}(\mathbf{Z}_1^k = \int dy_1 \cdots \int dy_M\, y_l \prod_{i=1}^{M} g(y_i) \prod_{l'=1}^{k} g\left(Z_{l'} - \sum_{j=1}^{M} A_j^{(l')} y_j\right), \tag{66}$$

$$I_2(\mathbf{Z}_1^k) = \int dy_1 \cdots \int dy_M \prod_{i=1}^{M} g(y_i) \prod_{l=1}^{k} g\left(Z_l - \sum_{j=1}^{M} A_j^{(l)} y_j\right). \tag{67}$$

As we shall see, these M-fold integrals can be explicitly evaluated in certain special cases.

For the figure of merit, we find

$$\epsilon_k^{*2} = E[(X_n - f_k^*)(X_n - f_k^*)] = E[(X_n - f_k^*)X_n]$$
$$= EX_n^2 - EX_n f_k^*, \tag{68}$$

since the best predictor $f_k^* = E(X_n|\mathbf{X}_{n-k}^{n-1})$ is uncorrelated with the prediction error $X_n - f_k^*$. For the two terms in (68), we have

$$EX_n^2 = \sum_{0}^{M} a_j^2 \tag{69}$$

from (38) and

$$EX_n f_k^* = -\sum_{j=1}^{k} S_{1j+1}^{(k+1)} EX_n X_{n-j} + \sum_{l=1}^{M} A_l^{(k+1)} E[X_n E(Y_{n-k-l}|\mathbf{X}_{n-k}^{n-1})] \tag{70}$$

from (47). These terms can be developed further as

$$J \equiv \sum_{j=1}^{k} S_{1j+1}^{(k+1)} EX_n X_{n-j} = \sum_{j=1}^{k} d_j \sum_{l=0}^{M-j} a_l a_{l+j}$$

$$= \sum_{\nu=1}^{M} a_\nu \sum_{l=(\nu-k)^+}^{\nu-1} a_l d_{\nu-l} \tag{71}$$

on letting $l + j = \nu$. But from (49)

$$\sum_{0}^{\nu-1} a_l d_{\nu-l} = -a_\nu, \quad \text{so if} \quad k \geq M,$$

$$J = \sum_{\nu=1}^{M} a_\nu \sum_{l=0}^{\nu-1} a_l d_{\nu-l} = -\sum_{1}^{M} a_\nu^2, \qquad k \geq M. \tag{72}$$

If $M > k$, (71) becomes

$$J = \sum_{\nu=1}^{k} a_\nu \sum_{l=0}^{\nu-1} a_l d_{\nu-l} + \sum_{\nu=k+1}^{M} a_\nu \sum_{l=\nu-k}^{\nu-1} a_l d_{\nu-l}$$

$$= -\sum_{\nu=1}^{k} a_\nu^2 + \sum_{\nu=1}^{M-k} a_{k+\nu} \sum_{l=0}^{k-1} a_{l+\nu} d_{k-l}. \tag{73}$$

For the last term in (70), we find

$$E[X_n E(Y_{n-k-l} | \mathbf{X}_{n-k}^{n-1})] = E\left[\left(\sum_{1}^{M} a_j Y_{n-j}\right) E(Y_{n-k-l} | \mathbf{Z}_1^k)\right].$$

Combining these results, we have

$$\epsilon_k^{*2} = 1 - \sum_{l=1}^{M} A_l^{(k+1)} \sum_{j=1}^{M} a_j E[Y_{n-j} E(Y_{n-k-l} | \mathbf{Z}_1^k)], \qquad k \geq M,$$

$$= \sum_{j=0}^{k} d_j \sum_{l=0}^{M-j} a_l a_{l+j}$$

$$- \sum_{j,l=1}^{M} A_l^{(k+1)} a_j E[Y_{n-j} E(Y_{n-k-l} | \mathbf{Z}_1^k)], \qquad k < M. \tag{74}$$

IV. THE SPECIAL CASE $M = 1$

When $M = 1$, we write $a_1 = -a$ so that

$$X_n = Y_n - aY_{n-1}. \tag{75}$$

From (48), $d_l = a^l$, and from (52), $A_1^{(l)} = -a^l$ so that (58) now reads

$$Z_{k+1-i} = \sum_{j=i}^{k} a^{j-i} X_{n-j} = Y_{n-i} - a^{k+1-i} Y_{n-(k+1)},$$

$$i = 1, 2, \cdots, k. \tag{76}$$

Equations (66) and (67) are

$$I_1 = \int dy\, yg(y) \prod_{l=1}^{k} g(Z_l + a^l y) \tag{77}$$

and

$$I_2 = \int dy\, g(y) \prod_{l=1}^{k} g(Z_l + a^l y). \tag{78}$$

For the best predictor, (47) now reads

$$f_k^*(\mathbf{X}_{n-k}^{n-1}) = -\sum_{j=1}^{k} a^{j-1} X_{n-j} - a^{k+1} I_1/I_2 \tag{79}$$

while (74) is

$$\epsilon_k^{*2} = 1 - a^{k+2} E[\, Y_{n-1} E(Y_{n-(k+1)} \,|\, \mathbf{Z}_1^k)]. \tag{80}$$

As an aid to the evaluation of (77) and (78), suppose that g has support from $y = \lambda$ to $y = \mu > \lambda$ so that $g(y) = 0$, $y < \lambda$, $y > \mu$. The integrands in (77) and (78) then vanish unless simultaneously $\lambda \leq y \leq \mu$ and $\lambda - Z_l \leq a^l y \leq \mu - Z_l$, $l = 1, 2, \cdots, k$. Thus, the integration in (77) and (78) can be restricted to the range

$$l \leq y \leq m,$$

$$l \equiv \max\left(\lambda, \frac{\lambda - Z_1}{a}, \frac{\lambda - Z_2}{a^2}, \cdots, \frac{\lambda - Z_k}{a^k}\right),$$

$$m \equiv \min\left(\mu, \frac{\mu - Z_1}{a}, \frac{\mu - Z_2}{a^2}, \cdots, \frac{\mu - Z_k}{a^k}\right), \qquad a > 0 \tag{81}$$

or, if $a = -b < 0$,

$$\bar{l} \leq y \leq \bar{m},$$

$$\bar{l} = \max\left(\lambda, \frac{Z_1 - \mu}{b}, \frac{\lambda - Z_2}{b^2}, \frac{Z_3 - \mu}{b^3}, \cdots\right)_{k+1},$$

$$\bar{m} = \min\left(\mu, \frac{Z_1 - \lambda}{b}, \frac{\mu - Z_2}{b^2}, \frac{Z_3 - \lambda}{b^3}, \cdots\right)_{k+1}, \tag{82}$$

where there are $k + 1$ quantities within the parentheses. Now use (76) and the notational aid

$$\hat{Y}_j \equiv Y_{n-j}, \qquad j = 1, 2, \cdots. \tag{83}$$

Elementary operations show that

$$l = \hat{Y}_{k+1} + \frac{1}{a^k} \max_{j=1,\cdots,k+1} [a^{j-1}(\lambda - \hat{Y}_j)]$$

$$m = \hat{Y}_{k+1} + \frac{1}{a^k} \min_{j=1,\cdots,k+1} [a^{j-1}(\mu - \hat{Y}_j)]$$

$$\bar{l} = \hat{Y}_{k+1} + \frac{1}{b^k}$$
$$\cdot \max[b^k(\lambda - \hat{Y}_{k+1}),\ b^{k-1}(\hat{Y}_k - \mu),\ b^{k-2}(\lambda - \hat{Y}_{k-1}),\ \cdots]_{k+1}$$

$$\bar{m} = \hat{Y}_{k+1} + \frac{1}{b^k}$$
$$\cdot \min[b^k(\mu - \hat{Y}_{k+1}),\ b^{k-1}(\hat{Y}_k - \lambda),\ b^{k-2}(\mu - \hat{Y}_{k-1}),\ \cdots]_{k+1}. \quad (84)$$

Since $\lambda - \hat{Y} \leq 0$ and $\mu - \hat{Y} \geq 0$, we see that $l \leq m$ and $\bar{l} \leq \bar{m}$ with probability one.

4.1 Y's are uniform

Let Y_j have the density

$$p_Y(y) = g(y) = \begin{cases} \dfrac{1}{2\gamma}, & |y| \leq \gamma \\ 0, & |y| > \gamma \end{cases}$$

$$\gamma \equiv \sqrt{3}. \quad (85)$$

Now $\lambda = -\gamma$, $\mu = \gamma$,

$$I_1 = \int_l^m dy \left(\frac{1}{2\gamma}\right)^{k+1} y = \left(\frac{1}{2\gamma}\right)^{k+1} \frac{1}{2}(m^2 - l^2),$$

$$I_2 = \int_l^m dy \left(\frac{1}{2\gamma}\right)^{k+1} = \left(\frac{1}{2\gamma}\right)^{k+1} (m - l),$$

and from (79)

$$f_k^*(\mathbf{X}_{n-k}^{n-1}) = -\sum_1^k a^{j-1} X_{n-j} - \frac{a^{k+1}}{2}(l + m), \quad (86)$$

while (80) is

$$\epsilon_k^{*2} = 1 - \frac{a^{k+2}}{2} E[Y_{n-1}(l - m)]. \quad (87)$$

It is a matter of straightforward algebra to put (86), (81), and (76) in the form (17) to (18). We omit the details here.

The evaluation of (87) can proceed as follows. Since $\hat{Y}_{k+1} = Y_{n-(k+1)}$ and $Y_{n-1} = \hat{Y}_1$ are independent, from (84)

$$\begin{aligned}\tfrac{1}{2}a^k E[Y_{n-1}(l+m)] &= \tfrac{1}{2}E[\hat{Y}_1 \max_{j=1,\cdots,k+1}(a^{j-1}(-\gamma - \hat{Y}_j))] \\ &\quad + \tfrac{1}{2}E[\hat{Y}_1 \min_{j=1,\cdots,k+1}(a^{j-1}(\gamma - \hat{Y}_j))] \\ &= 6E[(1-2\delta_0)\min_{j=0,\cdots,k}(\delta_j a^j)]. \end{aligned} \tag{88}$$

Here we have set $\hat{Y}_j = \gamma(1 - 2\delta_{j-1})$, $j = 1, 2, \cdots, k+1$, so that the δ's are i.i.d. variates uniform on (0, 1). Then from (87) and (88),

$$\epsilon_k^{*2} = 1 - 6a^2 E(1-2\delta)\min(\delta, \bar{\delta}) \tag{89}$$

where $\delta = \delta_0$ and

$$\bar{\delta} = \min_{1 \le j \le k}(\delta_j a^j). \tag{90}$$

Note that, since δ and $\bar{\delta}$ are independent,

$$\begin{aligned} &E(1-2\delta)\min(\delta, \bar{\delta}) \\ &\quad = \int_0^1 (1-2x)x\Pr(\delta \in dx, \bar{\delta} > x) \\ &\qquad + \int_0^1 \int_0^x (1-2x)y\Pr(\delta \in dx, \bar{\delta} \in dy) \\ &\quad = \int_0^1 (1-2x)x\Pr(\bar{\delta} > x)\, dx \\ &\qquad + \int_0^1 \int_0^x (1-2x)y \frac{d}{dy}[-\Pr(y < \bar{\delta} < x)]\, dy\, dx. \end{aligned} \tag{91}$$

Integrate by parts in the integral on dy. The boundary terms vanish and so

$$\begin{aligned} E(1-2\delta)\min(\delta, \bar{\delta}) &= \int_0^1 (1-2x)x\Pr(\bar{\delta} > x)\, dx \\ &\quad + \int_0^1 \int_0^x (1-2x)\Pr(y < \bar{\delta} < x)\, dy\, dx \\ &= \int_0^1 (1-2x)x\Pr(\bar{\delta} > x)\, dx \end{aligned}$$

$$+\int_0^1\int_0^x (1-2x)$$
$$\cdot[\Pr(\bar{\delta}>y)-\Pr(\bar{\delta}>x)]\,dy\,dx$$
$$=\int_0^1\int_0^x (1-2x)\Pr(\bar{\delta}>y)\,dy\,dx$$
$$=\int_0^1\int_y^1 (1-2x)\Pr(\bar{\delta}>y)\,dx\,dy$$
$$=-\int_0^1 y(1-y)\Pr(\bar{\delta}>y)\,dy. \tag{92}$$

Since from (90),

$$\Pr(\bar{\delta}>y)=\prod_{j=1}^{k}\left(1-\frac{y}{a^j}\right)^+ \tag{93}$$

(20) follows immediately from (89) and (92).

Equation (21) follows from (20) by setting

$$P_k(x)\equiv\int_0^1 du\,u(1-u)\prod_{i=1}^{k}(1-ux^i). \tag{94}$$

Now let

$$R_k(x,u)\equiv\prod_{i=1}^{k}(1-ux^i)=\sum_{j=0}^{k}b_ju^j \tag{95}$$

where the b's depend on k and x. Substitution in (94) yields the first part of (23) since $\int_0^1 du\,u(1-u)u^l=1/(l+2)(l+3)$. However, one sees from the product form for R_k in (95) that

$$(1-ux)R_k(x,ux)=(1-ux^{k+1})R_k(x,u)$$

so that

$$(1-ux)\sum_0^k b_j(ux)^j=(1-ux^{k+1})\sum_0^k b_ju^j.$$

Equating the coefficient of u^l on both sides of this equation yields

$$b_0=1,\quad b_l(1-x^l)=(x^k-x^l)b_{l-1},\quad l=1,2,\cdots,k$$

from which (22) and (23) then follow directly.

4.2 Y's are exponential

Let Y_j have the density

$$p_Y(y) = \begin{cases} e^{-(y+1)}, & y \geq -1 \\ 0, & y < -1. \end{cases} \tag{96}$$

Now $\lambda = -1$ and $\mu = \infty$ so that (81) is

$$l = -\min\left(1, \frac{1+Z_1}{a}\, \frac{1+Z_2}{a^2}, \cdots, \frac{1+Z_k}{a^k}\right),$$
$$m = \infty \qquad a > 0 \tag{97}$$

while

$$\bar{l} = -\min\left(1, \frac{1+Z_2}{b^2}, \frac{1+Z_4}{b^4}, \cdots\right)$$
$$\bar{m} = \min\left(\frac{1+Z_1}{b}, \frac{1+Z_3}{b^3}, \frac{1+Z_5}{b^5}, \cdots\right). \qquad a = -b < 0 \tag{98}$$

From (65), (77), and (78), we then find

$$E(Y_{n-(k+1)} | \mathbf{X}_{n-k}^{n-1}) = I_1/I_2$$
$$= \begin{cases} l + \dfrac{1}{B} & a > 0 \\ \dfrac{1}{B} + \dfrac{\bar{m}\, e^{-B\bar{m}} - \bar{l} e^{-B\bar{l}}}{e^{-B\bar{m}} - e^{-B\bar{l}}}, & a = -b < 0 \end{cases}$$

$$B \equiv 1 + a + a^2 + \cdots + a^k = \frac{1 - a^{k+1}}{1 - a}. \tag{99}$$

Using (97) to (99), it is a matter of straightforward algebra to convert (79) into the results stated as (25), (26), and (28). Along the way, use must be made of (76) which defines the Z's in terms of the observed X's. We omit the details here.

The evaluation of ϵ_k^{*2} is somewhat more complicated. Equations (80), (59), (83), and (99) give

$$\epsilon_k^{*2} = 1 - a^{k+2} E(\hat{Y}_1 I_1/I_2) \tag{100}$$

and from (99) it is seen that the cases $a > 0$ and $a < 0$ must be treated separately.

When $a > 0$, $I_1/I_2 = l + (1/B)$. Then (100) becomes

$$\epsilon_k^{*2} = 1 - a^{k+2} E\left[\hat{Y}_1 \left(\frac{1}{B} + \hat{Y}_{k+1} + \frac{1}{a^k} \max_{j=1,\ldots,k+1} [a^{j-1}(-1 - \hat{Y}_j)]\right)\right]$$

$$= 1 + a^2E(\theta_1 - 1)\min[\theta_1, a\theta_2, \cdots, a^k\theta_{k+1}], \tag{101}$$

$$\theta_j \equiv 1 + \hat{Y}_j, \quad j = 1, 2, \cdots, k+1. \tag{102}$$

Here we have used (84) to express l in terms of the $\hat{Y}$'s and have used the fact that $E\hat{Y}_1 = 0$ and that the $\hat{Y}$'s are independent. Now define

$$U \equiv \min(a\theta_2, a^2\theta_3, \cdots, a^k\theta_{k+1}).$$

Since each θ has the one-sided exponential distribution

$$p_{\theta_1}(\ ) = \begin{cases} e^{-\theta}, & \theta \geq 0 \\ 0, & \theta < 0 \end{cases} \tag{103}$$

one readily finds that the density for U is

$$p_U(u) = \begin{cases} \frac{1}{\delta} e^{-u/\delta}, & u > 0 \\ 0, & u < 0 \end{cases}$$

with δ as in (26). Furthermore, U and θ_1 are independent. The calculation of (101) then reads

$$\epsilon_k^{*2} = 1 + a^2E(\theta_1 - 1)\min(\theta_1, U)$$

$$= 1 + a^2\left[\int_0^\infty d\theta(\theta - 1)\int_0^\infty du\, up_U(u)p_{\theta_1}(\theta)\right.$$

$$\left. + \int_0^\infty d\theta(\theta - 1)\theta\int_\theta^\infty du\, p_U(u)p_{\theta_1}(\theta)\right].$$

The integrals are readily evaluated and (27) results.

The computation of ϵ_k^{*2} is more burdensome when $a = -b < 0$. From (99) and (100),

$$\epsilon_k^{*2} = 1 - (-1)^k b^{k+2} E\left[\hat{Y}_1 \frac{\bar{m}e^{-B\bar{m}} - \bar{l}e^{-B\bar{l}}}{e^{-B\bar{m}} - e^{-B\bar{l}}}\right], \tag{104}$$

where from (84)

$$\bar{l} = \hat{Y}_{k+1} + \frac{1}{b^k}\max[-b^k\theta_{k+1}, -b^{k-2}\theta_{k-1}, \cdots]$$

$$\bar{m} = \hat{Y}_{k+1} + \frac{1}{b^k}\min[b^{k-1}\theta_k, b^{k-3}\theta_{k-2}, \cdots]$$

and we have again used (102). Now write

$$\bar{B} \equiv B/b^k, \qquad \bar{l} = \hat{Y}_{k+1} - b^{-k}\bar{U}, \qquad \bar{m} = \hat{Y}_{k+1} + b^{-k}\bar{V}.$$

Now (104) becomes

$$\epsilon_k^{*2} = 1 - (-1)^k b^2 E(\theta_1 - 1)\frac{\bar{V}e^{-\bar{B}\bar{V}} + \bar{U}e^{\bar{B}\bar{U}}}{e^{-\bar{B}\bar{V}} - e^{\bar{B}\bar{U}}}, \tag{105}$$

$$\bar{U} = \min[b^k\theta_{k+1}, b^{k-2}\theta_{k-1}, \cdots],$$

$$\bar{V} = \min[b^{k-1}\theta_k, b^{k-3}\theta_{k-2}, \cdots], \tag{106}$$

where $\theta_1, \theta_2, \cdots, \theta_{k+1}$ are i.i.d. random variables having the density (103).

Suppose now that $k = 2l$ is even. We have

$$\bar{U} = \min(\theta_1, Z)$$

$$\bar{Z} = \min(b^2\theta_3, b^4\theta_5, \cdots, b^{2l}\theta_{2l+1})$$

$$\bar{V} = \min(b\theta_2, b^3\theta_4, \cdots, b^{2l-1}\theta_{2l})$$

and

$$p_Z(z) = \begin{cases} B_e e^{-B_e z}, & z \geq 0 \\ 0, & z < 0 \end{cases}, \qquad B_e = \frac{1}{b^2} + \frac{1}{b^4} + \cdots + \frac{1}{b^{2l}}$$

$$p_V(v) = \begin{cases} B_0 e^{-B_0 v}, & v \geq 0 \\ 0, & v < 0 \end{cases}, \qquad B_0 = \frac{1}{b} + \frac{1}{b^3} + \cdots + \frac{1}{b^{2l-1}}.$$

Since θ_1, Z, and $\bar{V}$ are independent random variables,

$$\epsilon_k^{*2} = 1 - b^2 \int_0^\infty dv\, p_V(v)$$

$$\cdot\left[\int_0^\infty d\theta\, p_{\theta_1}(\theta)(\theta - 1)\int_\theta^\infty dz\, p_Z(z)\frac{ve^{-\bar{B}v} + \theta e^{\bar{B}\theta}}{e^{-\bar{B}v} - e^{\bar{B}\theta}}\right.$$

$$\left. + \int_0^\infty d\theta\, p_{\theta_1}(\theta)(\theta - 1)\int_0^\theta dz\, p_Z(z)\frac{ve^{-\bar{B}v} + ye^{\bar{B}z}}{e^{-\bar{B}v} - e^{\bar{B}z}}\right].$$

Inside the brackets here, the z integration can be carried out immediately in the first term. Interchange order of integration of θ and z in

the second term and carry out tbe θ integration. There results

$$\epsilon_k^{*2} = 1 - b^2 B_0 \int_0^\infty dv \int_0^\infty d\theta\, e^{-\bar{B}\theta} e^{-B_0(\theta+v)} \frac{v e^{-\bar{B}(v+\theta)} + \theta}{e^{-\bar{B}(v+\theta)} - 1} [(B_e + 1)\theta - 1].$$

Now change variables of integration to x and y via $v + \theta = x$, $\theta = y$. Tedious, but straightforward integration now gives

$$\epsilon_k^{*2} = 1 + b^2 \left[\frac{B_0^2 + \bar{B}^2}{\bar{B}^2 (B_0 + \bar{B})^2} - S \right] \tag{107}$$

where

$$S = \frac{B_0(B_e + 1)}{\bar{B}} \int_0^\infty dx \frac{x^2 e^{-(B_0 + 2\bar{B})x}}{1 - e^{-\bar{B}x}}. \tag{108}$$

Now

$$\bar{B} = B/b^k = \frac{1 - b^{k+1}(-1)^{k+1}}{b^k(1 + b)} \tag{109}$$

from (99). Since k is even in the present case, $\bar{B} > 0$ and the factor $[1 - e^{-\bar{B}x}]^{-1}$ in (108) can be expanded as $\sum_0^\infty e^{-n\bar{B}x}$. Term-by-term integration, insertion in (107), and a little rearranging yield (29) for k even.

The case of odd k proceeds in a similar manner. Now $\bar{B}$ as given by (109) will be negative if $b > 1$. An integral of the form (108) that occurs in the calculation must now be expanded in two different ways depending on whether $\bar{B}$ is positive or negative. This gives rise to the several forms for λ in (29). We omit the straghtforward details of calculation here.

4.3 Y's are discrete binary

Now let Y take only two values:

$$\Pr[Y = \lambda] = p, \quad \Pr[Y = \mu] = q \equiv 1 - p$$

$$\lambda = -\sqrt{\frac{q}{p}}, \mu = \sqrt{\frac{p}{q}}$$

so that as always $EY = 0$, $EY^2 = 1$. When $k = 1$ and $a_1 = -1$, the $d_l = 1, A_1^{(l)} = -1$ and (38), (47), and (58) become

$$\begin{aligned} X_n &= Y_n - Y_{n-1} \\ f_k^* &= -Z_k - E(Y_{n-(k+1)} | Z_1, \cdots, Z_k) \\ Z_k &= Y_{n-1} - Y_{n-(k+1)} \\ &\vdots \\ Z_1 &= Y_{n-k} - Y_{n-(k+1)}. \end{aligned} \tag{110}$$

Here each Z can take only values $0, \mu - \lambda, \lambda - \mu$. One readily finds

$$\Pr[Y_{n-(k+1)} = \lambda \mid \mathbf{Z}_1^k] = \begin{cases} 0, & \text{some } Z < 0 \\ 1, & \text{some } Z > 0 \\ \dfrac{p^{k+1}}{p^{k+1} + q^{k+1}}, & \text{all } Z\text{'s } 0 \end{cases}$$

$$\Pr[Y_{n-(k+1)} = \mu \mid \mathbf{Z}_1^k] = \begin{cases} 0, & \text{some } Z > 0 \\ 1, & \text{some } Z < 0 \\ \dfrac{q^{k+1}}{p^{k+1} + q^{k+1}}, & \text{all } Z\text{'s } 0 \end{cases}$$

so that

$$E(Y_{n-(k+1)} \mid \mathbf{Z}_1^k) = \begin{cases} \lambda, & \text{at least one } Z > 0 \\ \mu, & \text{at least one } Z < 0 \\ \dfrac{\lambda p^{k+1} + \mu q^{k+1}}{p^{k+1} + q^{k+1}}, & \text{all } Z\text{'s zero.} \end{cases} \tag{111}$$

Equation (32) follows then from (110) and (111).

For the figure of merit of the best predictor, we have in this case

$$\begin{aligned} \epsilon_k^{*2} &= 2 - EX_n f_k^* \\ &= 2 - E\{[Y_n - Y_{n-1}] \\ &\qquad \cdot [-Y_{n-1} + Y_{n-(k+1)} - E(Y_{n-(k+1)} \mid \mathbf{Z}_1^k)]\} \\ &= 1 - E[Y_{n-1} E(Y_{n-(k+1)} \mid \mathbf{Z}_1^k)]. \end{aligned} \tag{112}$$

Now

$$\begin{aligned} EY_{n-1}E(Y_{n-(k+1)} \mid \mathbf{Z}_1^k) = \lambda &\frac{\lambda p^{k+1} + \mu q^{k+1}}{p^{k+1} + q^{k+1}} \\ &\cdot \Pr[Y_{n-1} = \lambda, \quad \text{all } Z\text{'s} = 0] \\ &+ \mu \frac{\lambda p^{k+1} + \mu q^{k+1}}{p^{k+1} + q^{k+1}} \\ &\cdot \Pr[Y_{n-1} = \mu, \quad \text{all } Z\text{'s}, = 0] \\ &+ \lambda^2 \Pr[Y_{n-1} = \lambda, \quad \text{some } Z > 0] \\ &+ \lambda\mu \Pr[Y_{n-1} = \mu, \quad \text{some } Z > 0] \\ &+ \lambda\mu \Pr[Y_{n-1} = \lambda, \quad \text{some } Z < 0] \\ &+ \mu^2 \Pr[Y_{n-1} = \mu, \quad \text{some } Z < 0]. \end{aligned} \tag{113}$$

The six probabilities $\Pr[\ ]$ listed here are readily seen to be p^{k+1}, q^{k+1},

$p^2[1-p^{k+1}]$, qp, pq, and $q^2[1-q^{k+1}]$, respectively. Equation (35) then follows from (112) and (113) by simple algebra.

When $a=-1$, the corresponding equations are

$$\begin{aligned}
X_n &= Y_n + Y_{n-1} \\
Z_k &= Y_{n-1} - (-1)^k Y_{n-(k+1)} \\
Z_{k-1} &= Y_{n-2} - (-1)^{k+1} Y_{n-(k+1)} \\
&\vdots \\
Z_1 &= Y_{n-k} - (-1)^1 Y_{n-(k+1)} \\
f_k^* &= Z_k - (-1)^{k+1} E(Y_{n-(k+1)} | \mathbf{Z}_1^k) \\
\epsilon_k^{*2} &= 1 - (-1)^k E[Y_{n-1} E(Y_{n-(k+1)} | \mathbf{Z}_1^k)].
\end{aligned}$$

Note that the Z's with odd subscript can take values $(\lambda+\mu)$, 2λ, or 2μ while the Z's with even subscript can take values $\lambda-\mu$, 0 or $\mu-\lambda$. Let there be s Z's with even subscripts and t Z's with odd subscripts. Then

$$\Pr[Y_{n-(k+1)} = \lambda | \mathbf{Z}_1^k] = \begin{cases} \dfrac{p^{s+1}q^t}{p^{s+1}q^t + q^{s+1}p^t}, & \text{all } Z\text{'s} = (\lambda+\mu) \text{ or zero} \\ 1, & \text{some } Z = 2\lambda \text{ or } \mu-\lambda, \end{cases}$$

$$\Pr[Y_{n-(k+1)} = \mu | \mathbf{Z}_1^k] = \begin{cases} \dfrac{q^{s+1}p^t}{p^{s+1}q^t + q^{s+1}p^t}, & \text{all } Z\text{'s} = \lambda+\mu \text{ or } 0 \\ 1, & \text{some } Z = 2\mu \text{ or } \lambda-\mu \end{cases}$$

and it follows that

$$E(Y_{n-(k+1)} | \mathbf{Z}_1^k) = \begin{cases} \dfrac{\lambda p^{s+1}q^t + \mu q^{s+1}p^t}{p^{s+1}q^t + q^{s+1}p^t}, & \text{all } Z\text{'s} = \lambda+\mu \text{ or zero} \\ \lambda, & \text{some } Z = 2\lambda \text{ or } \mu-\lambda \\ \mu, & \text{some } Z = 2\mu \text{ or } \lambda-\mu. \end{cases}$$

Equation (33) then follows at once. The computation of ϵ_k^{*2} is now a

little more complicated in that the cases k even and k odd must be treated separately. We list the key equation in the computation.

$$E[Y_{n-1}E(Y_{n-(k+1)}|\mathbf{Z}_1^k)] = \frac{\lambda p^{s+1}q^t + \mu q^{s+1}p^t}{p^{s+1}q^t + q^{s+1}p^t}$$

$$\cdot\left[\lambda\begin{Bmatrix}p^{s+1}q^t\\ p^tq^{s+1}\end{Bmatrix} + \mu\begin{Bmatrix}q^{s+1}p^t\\ q^tp^{s+1}\end{Bmatrix}\right.$$

$$+ \lambda^2\begin{Bmatrix}p^2(1-q^tp^{s+1})\\ p^2\end{Bmatrix}$$

$$+ \lambda\mu\begin{Bmatrix}2pq\\ pq[(1-q^{t-1}p^s)+(1-p^{t-1}q^s)]\end{Bmatrix}$$

$$\left.+ \mu^2\begin{Bmatrix}q^2(1-p^tq^{s+1})\\ q^2\end{Bmatrix}\right.$$

Here the upper choice corresponds to k even and the lower choice to k odd. Straightforward manipulations now lead to (35) and (36). We omit the tedious details here.

V. THE CASE $a_j = \alpha^j$, $j = 1, 2, \cdots, M$

We now consider the exponential filter

$$X_n = \sum_{j=0}^{M} \alpha^j Y_{n-j}, \quad M > 1. \tag{114}$$

For the parameters of Section III, we have $a_j = \alpha^j, j = 0, 1, \cdots, M$, and all other a_j are zero. Equations (49) or (48) then yield

$$\begin{aligned} d_{j(M+1)} &= \alpha^{j(M+1)}\\ d_{j(M+1)+1} &= -\alpha^{j(M+1)+1}\\ d_{j(M+1)+1+s} &= 0\\ j &= 0, 1, 2, \cdots\\ s &= 1, 2, \cdots, M-1. \end{aligned} \tag{115}$$

From (52) we then find

$$\begin{aligned} A_l^{j(M+1)+1} &= \alpha^{j(M+1)+l}\\ A_l^{j(M+1)+1+s} &= \begin{cases}0, & l \neq M+1-s\\ -\alpha^{(j+1)(M+1)}, & l = M+1-s\end{cases}\\ j &= 0, 1, 2, \cdots\\ l &= 1, 2, \cdots, M\\ s &= 1, 2, \cdots, M. \end{aligned} \tag{116}$$

As before, we suppose that the density $p_{Y_i}(y) = g(y)$ has support on $\lambda \leq y \leq \mu$. To evaluate (66) and (67) it is necessary to determine the M-dimensional region of support for the integrands. The hyperplane-boundary constraints on $y_1 \cdots y_M$ are

$$\lambda \leq y_i \leq \mu, \quad i = 1, 2, \cdots, M \tag{117}$$

$$\lambda \leq Z_l - \sum_{j=1}^{M} A_j^{(l)} y_j \leq \mu \tag{118}$$

$$l = 1, 2, \cdots, k.$$

Now let

$$k = p(M + 1) + 1 + s \tag{119}$$

for some $p = 0, 1, 2, \cdots$ and some integer s such that $0 \leq s \leq M$. In view of (116), the constraints (118) on $y_1 \cdots y_M$ become

$$\lambda \leq Z_{\sigma(M+1)+1} - \alpha^{\sigma(M+1)} \sum_{l=1}^{M} \alpha^l y_l \leq \mu, \quad \sigma = 0, 1, \cdots, p$$

$$\lambda \leq Z_{\sigma(M+1)+1+j} + \alpha^{(\sigma+1)(M+1)} y_{M+1-j} \leq \mu$$

$$\sigma = 0, 1, \cdots, p-1, \quad j = 1, 2, \cdots, M$$

$$\lambda \leq Z_{p(M+1)+1+j} + \alpha^{(p+1)(M+1)} y_{M+1-j} \leq \mu$$

$$j = 1, 2, \cdots, s.$$

Notice that most of the hyperplane boundaries are parallel to the coordinate planes and that the remaining ones are all parallel. This is because of the simple exponential form of the filter (114). Thus we find the support for the integrands of (66) and (67) to be given by

$$l_j \leq y_j \leq m_j, \quad j = 1, 2, \cdots, M$$

$$A \leq \sum_{1}^{M} \alpha^j Y_j \leq B \tag{120}$$

where

$$A \equiv \max_{\sigma=0,1,\cdots,p} \left[\frac{Z_{\sigma(M+1)+1} - \mu}{\alpha^{\sigma(M+1)}} \right]$$

$$B = \min_{\sigma=0,1,\cdots,p} \left[\frac{Z_{\sigma(M+1)+1} - \lambda}{\alpha^{\sigma(M+1)}} \right]$$

$$l_j = \max \left[\lambda, \frac{\lambda - Z_{(M+1)+1-j}}{\alpha^{M+1}}, \frac{\lambda - Z_{2(M+1)+1-j}}{\alpha^{2(M+1)}} \right.$$

$$\cdots, \frac{\lambda - Z_{t(M+1)+1-j}}{\alpha^{t(M+1)}}\Bigg]$$

$$m_j = \min\Bigg[\mu, \frac{\mu - Z_{(M+1)+1-j}}{\alpha^{M+1}}, \frac{\mu - Z_{2(M+1)+1-j}}{\alpha^{2(M+1)}},$$

$$\cdots, \frac{\mu - Z_{t(M+1)+1-j}}{\alpha^{t(M+1)}}\Bigg]$$

$$t = \begin{cases} p, & j = 1, 2, \cdots, M - s \\ p + 1, & j = M - s + 1, \cdots, M. \end{cases} \tag{121}$$

The quantities A, B, l_j, m_j, for $j = 1, \cdots, M$ are random variables, and via (58) can be expressed in terms of the Y's. Equations (39) read in the present case

$$\begin{aligned} Z_{\sigma(M+1)+1-j} &= \hat{Y}_{\sigma(M+1)-j} - \alpha^{\sigma(M+1)}\hat{Y}_{-j} \\ & j = 1, 2, \cdots, M, \quad \sigma = 1, 2, \cdots, p \\ Z_{(p+1)(M+1)+1-j} &= \hat{Y}_{(p+1)(M+1)-j} - \alpha^{(p+1)(M+1)}\hat{Y}_{-j} \\ & j = M - s + 1, M - s + 2, \cdots, M \\ Z_{\sigma(M+1)+1} &= \hat{Y}_{\sigma(M+1)} + \alpha^{\sigma(M+1)} \sum_{l=1}^{M} \alpha^l \hat{Y}_{-l} \\ & \sigma = 0, 1, \cdots, p \end{aligned} \tag{122}$$

where for notational convenience we have put

$$\hat{Y}_u = Y_{n-k+u}, \quad u = 0, \pm 1, \pm 2, \cdots. \tag{123}$$

Equations (121) now become

$$A = \sum_{l=1}^{M} \alpha^l \hat{Y}_{-l} + \max_{\sigma=0,1,\cdots,p} \left[\frac{\hat{Y}_{\sigma(M+1)} - \mu}{\alpha^{\sigma(M+1)}}\right]$$

$$B = \sum_{l=1}^{M} \alpha^l \hat{Y}_{-l} + \min_{\sigma=0,1,\cdots,p} \left[\frac{\hat{Y}_{\sigma(M+1)} - \lambda}{\alpha^{\sigma(M+1)}}\right].$$

$$l_j = \hat{Y}_{-j} + \max\left[\lambda - \hat{Y}_{-j}, \frac{\lambda - \hat{Y}_{\sigma(M+1)-j}}{\alpha^{\sigma(M+1)}}; \quad \sigma = 1, 2, \cdots t\right]$$

$$m_j = Y_{-j} + \min\left[\mu - \hat{Y}_{-j}, \frac{\mu - \hat{Y}_{\sigma(M+1)-j}}{\alpha^{\sigma(M+1)}}; \quad \sigma = 1, 2, \cdots, t\right]$$

$$t = \begin{cases} p, & j = 1, 2, \cdots, M - s \\ p + 1, & j = M - s + 1, M - s + 2, \cdots, M. \end{cases} \tag{124}$$

5.1 Y's exponential

Let the Y's be i.i.d. with density (96). Now $\lambda = -1$, $\mu = \infty$. Then (67) can be written

$$I_2 = c \int_{l_1}^{\infty} dy_1 \cdots \int_{l_M}^{\infty} dy_M \, e^{-\sum_1^M y_i + \sum_{i=1}^{k} \sum_1^M A_j^{(i)} y_j}$$

$$\sum_{j=1}^{M} \alpha^j y_j \leq B. \tag{125}$$

I_{l1} will be given by a similar expression with an extra factor of y_l in the integrand.

It is convenient now to set $x_j = (y_j - l_j)\alpha^j$ to obtain

$$I_2 = \hat{c} \int_0^{\infty} dx_1 \cdots \int_0^{\infty} dx_M \, e^{\sum_1^M \beta_j x_j}$$

$$\sum_{j=1}^{M} x_j \leq \hat{B} \tag{126}$$

where

$$\hat{B} = B - \sum_1^M \alpha^j l_j \tag{127}$$

and

$$\beta_j = \alpha^{-j}\left[\sum_{i=1}^{k} A_j^{(i)} - 1\right] = \frac{1 - \alpha^{(p+1)(M+1)}}{1 - \alpha^{M+1}} - \frac{\alpha^{-j}}{1 - \alpha^{M+1}}$$

$$\times \begin{cases} 1 - \alpha^{(p+1)(M+1)}, & j = 1, 2, \cdots, M - s \\ 1 - \alpha^{(p+2)(M+1)}, & j = M - s + 1, \cdots, M. \end{cases} \tag{128}$$

But the integral in (126) can be easily evaluated. Denote $I_2/\hat{c}$ by $J_M(\beta_1, \cdots, \beta_M; \hat{B})$. Then integration on x_M yields the recurrence

$$J_M(\beta_1, \cdots, \beta_M; \hat{B}) = \frac{1}{\beta_M}$$

$$\cdot[e^{\hat{B}\beta_M} J_{M-1}(\beta_1 - \beta_M, \beta_2 - \beta_M, \cdots, \beta_{M-1} - \beta_M; \hat{B})$$

$$- J_{M-1}(\beta_1, \beta_2, \cdots, \beta_{M-1}; \hat{B})]$$

with $J_1(\beta_1; \hat{B}) = [e^{\hat{B}\beta_1} - 1]/\beta_1$. The solution is

$$J_m(\beta_1, \cdots, \beta_M; \hat{B}) = (-1)^M \left[\frac{1}{\prod_1^M \beta_j} - \sum_{j=1}^{M} \frac{e^{\hat{B}\beta_j}}{\beta_j \prod_{k \neq j}^{M} (\beta_k - \beta_j)} \right] \tag{129}$$

as can be shown by induction.

As noted, I_{l1} differs from (125) by a factor y_l in the integrand, or from (126) by a factor x_l/α^l in the integrand. From this, it is seen that $I_{l1} = \alpha^{-l}(\partial J_M/\partial \beta_l)$ so that from (65)

$$E(Y_{n-k-l} | \mathbf{X}_{n-k}^{n-1}) = \frac{1}{\alpha^l} \frac{\partial}{\partial \beta_l} \log J_M(\beta_1, \beta_2, \cdots, \beta_M; \hat{B}) \tag{130}$$

where J_M is given by (129) and the other parameters by (127) and (128). Expression (130) must be inserted into (47) to obtain the complete estimator. From (115) and (116) we see that the form of the best estimator depends on s defined in (119). Thus

$$f_k^*(\mathbf{X}_{n-k}^{n-1}) = -\sum_{j=1}^{p} \alpha^{j(M+1)}(X_{n-j(M+1)} - \alpha X_{n-j(M+1)-1})$$

$$+ \begin{cases} \alpha^{(p+1)(M+1)} \sum_{l=1}^{M} \alpha^l E(Y_{n-k-l} | \mathbf{X}_{n-k}^{n-1}), & s = M \\ -\alpha^{(p+1)(M+1)} E(Y_{n-k} - (M-s) | \mathbf{X}_{n-k}^{n-1}), & s = 0, 1, \cdots, M-1 \end{cases} \tag{131}$$

where the first sum is vacuous if $p = 0$.

When $s = M$, the last sum in (131) can be carried out. We emphasize how complicated the best estimator is for this relatively simple case by writing it out in full:

$$k = \nu(M+1)$$

$$f_k^*(\mathbf{X}_{n-k}^{n-1}) = -\sum_{j=1}^{\nu-1} \alpha^{j(M+1)}(X_{n-j(M+1)} - \alpha X_{n-j(M+1)-1})$$

$$+ \alpha^{\nu(M+1)} \hat{B} + (-1)^M H_M / J_M,$$

$$H_M = -\frac{\hat{B} + \sum \frac{1}{\beta_j}}{\prod \beta_j} + \sum_{l=1}^{M} \frac{e^{\hat{B}\beta_l}}{\beta_l^2 \prod_{k \neq l} (\beta_k - \beta_l)},$$

$$\beta_j = \frac{1 - \alpha^{\nu(M+1)} - \alpha^{-j}(1 - \alpha^{(\nu+1)(M+1)})}{1 - \alpha^{M+1}}. \tag{132}$$

Here J_M is given by (129) and the observable $\hat{B}$ is given by (127) with B and the l_j being given by (121). Finally, from (58) the Z's of these

equations are given in terms of the observable X's by

$$Z_i = \sum_{j=0}^{i-1} d_j X_{n-k-(j+1-i)}, \quad i = 1, 2, \cdots, k \tag{133}$$

with the d's as in (115).

The expression for ϵ_k^{*2} can be reduced to a two-dimensional integral, but there seems to be little point in exhibiting this exceedingly complicated expression.

5.2 Y's uniform

Explicit formulas for the best estimator f_k^* can be worked out when the Y's are i.i.d. with the density (85). Again, the results are exceedingly complicated.

We now have $\lambda = -\mu = \gamma = \sqrt{3}$ and (67) can be written

$$I_2 = c' \int_{l_1}^{m_1} dy_1 \cdots \int_{l_M}^{m_M} dy_M$$

$$A \leq \sum \alpha^l y_l \leq B \tag{134}$$

with the l's, m's, A and B given by (121). The quantity I_{l1} of (66) is similar to (134) except for a factor of y_l in the integrand.

To evaluate (134), let

$$\alpha^j y_j = x_j, \qquad \alpha^j m_j = \hat{m}_j, \qquad \alpha^j l_j = \hat{l}_j$$

so that

$$I_2 = c'' \int_{\hat{l}_1}^{\hat{m}_1} dx_1 \cdots \int_{\hat{l}_M}^{\hat{m}_M} dx_M.$$

$$A \leq \sum x_j \leq B. \tag{135}$$

We can now interpret $I_2/c'' \prod_1^M (\hat{m}_i - \hat{l}_i)$ as $\Pr[A \leq \sum_1^M T_i \leq B]$ where the T's are independent random variables and T_i is uniformly distributed between $\hat{l}_i$ and $\hat{m}_i$, $i = 1, 2, \cdots, M$. The characteristic function of the random variable $S \equiv \sum T_i$ is

$$\Phi_S(f) = \prod_1^M \Phi_{T_j}(f) = \prod_1^M \frac{e^{if\hat{m}_j} - e^{if\hat{l}_j}}{if(\hat{m}_j - \hat{l}_j)}$$

so that the density for S is

$$p_S(s) = \int_{-\infty}^{\infty} df\, e^{-isf} \Phi_S(f)$$

$$= K \int_{\mathcal{C}} df\, e^{-isf} \prod_{j=1}^M \frac{e^{if\hat{m}_j} - e^{if\hat{l}_j}}{f}$$

$$= K \sum_{jk} (-1)^j \int_{\mathscr{C}} \frac{e^{if(\Gamma_{jk}-s)}}{f^M},$$

$$K = 1/(2\pi i^M \prod (\hat{m}_j - \hat{l}_j)). \tag{136}$$

Here $\mathscr{C}$ is a contour in the complex plane that runs along the real axis except for a semicircular excursion into the negative imaginary half-plane along a circle with center at the origin. The 2^M quantities Γ_{jk} arise from multiplying out the product and are given by

$$\Gamma_{jk} = \hat{l}_{i_1} + \hat{l}_{i_2} + \cdots + \hat{l}_{i_j} + \hat{m}_{i_{j+1}} + \cdots \hat{m}_{i_M} \tag{137}$$

where $i_1, i_2, \cdots, i_M$ is a permutation of the integers $1, 2, \cdots, M$ and $i_1, i_2, \cdots, i_j$ are chosen in all $\binom{M}{j}$ ways. Thus

$$k = 1, 2, \cdots, \binom{M}{j}$$

$$j = 0, 1, \cdots, M. \tag{138}$$

Now

$$\int_{\mathscr{C}} \frac{e^{ifu}}{f^M}\, df = \begin{cases} \dfrac{2\pi i^M}{(M+1)!} u^{M-1}, & u > 0 \\ 0, & u \leq 0 \end{cases}$$

as can be readily established by contour integration, so that

$$p_S(s) = \frac{1}{(M-1)! \prod (\hat{m}_j - \hat{l}_j)} \sum_{j=0}^{M} (-1)^j \sum_{i=1}^{\binom{M}{j}} [(\Gamma_{jk} - s)^+]^{M-1}. \tag{139}$$

Thus since $\Pr[A \leq S \leq B] = \int_A^B p_S(s)\, ds$ we find finally

$$I_2 = \frac{c''}{M!} \sum_{j=0}^{M} (-1)^j \sum_{i=1}^{\binom{M}{j}} \{[(\Gamma_{jk} - A)^+]^M - [(\Gamma_{jk} - B)^+]^M\}. \tag{140}$$

For the numerator in (65), we have

$$I\nu_1/c'' \prod_1^M (\hat{m}_i - \hat{l}_i) = \int_{\hat{l}_\nu}^{\hat{m}_\nu} dt \int_A^B ds\, t p_{T_\nu S}(t, s) \tag{141}$$

with S and the T's as before. Now

$$\begin{aligned} p_{T_\nu S}(t, s) &= p_{T_\nu}(t) p_{S|T_\nu}(s \mid t) \\ &= p_{T_\nu}(t) p_{S_\nu}(s - t) \end{aligned} \tag{142}$$

where

$$S_\nu \equiv \sum_{j \neq \nu} T_j$$

is independent of T_ν. By the same steps that led to (139)

$$ps_\nu(s) = \frac{1}{(M-2)! \prod_{j \neq \nu} (\hat{m}_j - \hat{l}_j)} \sum_{j=0}^{M-1} (-1)^j \sum_{i=1}^{\binom{M-1}{j}} [(\Gamma_{jk}^{(\nu)} - s)^+]^{M-2} \quad (143)$$

where the $\Gamma_{jk}^{(\nu)}$ are formed in a manner analogous to Γ_{jk}, only $\hat{l}_\nu$ and $\hat{m}_\nu$ are omitted from (137). Inserting (143) and (142) into (141), one finds finally

$$\begin{aligned} I_{\nu 1} = \frac{c''}{(M+1)!} \sum_{j=0}^{M-1} (-1)^j \sum_{i=1}^{\binom{M-1}{j}} & \{[(\Gamma_{jk}^{\nu} - A + m_\nu)^+]^M \\ & \cdot [(M+1)m_\nu - (\Gamma_{jk}^{\nu} - A + m_\nu)^+] \\ & - [(\Gamma_{jk}^{\nu} - A + l_\nu)^+]^M [(M+1)l_\nu - (\Gamma_{jk}^{\nu} - A + l_\nu)^+] \\ & - [(\Gamma_{jk}^{\nu} - B + m_\nu)^+]^M [(M+1)m_\nu - (\Gamma_{jk}^{\nu} - B + m_\nu)^+] \\ & + [(\Gamma_{jk}^{\nu} - B + l_\nu)^+]^M [(M+1)l_\nu - (\Gamma_{jk}^{\nu} - B + l_\nu)^+]. \end{aligned} \quad (144)$$

The ratio of (141) to (140), which is independent of c'', is $E(Y_{n-k-\nu} | \mathbf{X}_{n-k}^{n-1})$. The best estimator is obtained by using this quantity in (131).

VI. ENTROPY INEQUALITY

We begin by giving some definitions and stating some facts concerning the Shannon differential entropy of a random variable. Let U be a real-valued random variable with probability density function $p_U(u)$, $-\infty < u < \infty$. The (*differential*) *entropy* of U is

$$H(U) = -\int_{-\infty}^{\infty} p_U(u) \log p_U(u)\, du. \quad (145)$$

Intuitively, $H(U)$ can be thought of as a measure of the spread of the density function $p_U(\cdot)$. The following facts are easily verifiable (see, for example, Refs. 3 and 4).

(*a*) $H(U)$ can take any value in $[-\infty, +\infty]$. However, for $s > 0$,

$$H(U) \leq \frac{1}{s} \log \frac{2^s e \Gamma^s(1/s) E|U|^s}{s^{s-1}}, \quad (146)$$

with equality when $p_U(u)$ is of the form $K_1 \exp\{-K_2 |u|^s\}$, $-\infty < u < \infty$. The constant K_2 is a parameter and K_1 is chosen so that $\int p_U(u)\, du = 1$. Thus $E|U|^s < \infty$, for some $s > 0$, implies $H(U) < \infty$.

(*b*) For constant a, $-\infty < a < \infty$,

$$H(U) = H(U + a). \tag{147}$$

Inequality (146), for $s = 2$, and (147) imply that

$$H(U) \leq \frac{1}{2}\log[2\pi e(\text{var } U)],$$

with equality when U is Gaussian.

(c) For $-\infty < a < \infty$,

$$H(aU) = \log|a| + H(U). \tag{148}$$

Next assume that U and V are real-valued random variablea with joint probability density $p_{UV}(u, v)$, and marginal densities $p_U(u)$, $p_V(v)$, respectively, $-\infty < u, v < \infty$. The conditional entropy of U given V is usually defined as

$$H(U|V) = -\int_{-\infty}^{\infty} p_V(v)\, dv \left\{\int_{-\infty}^{\infty} p_{U|V}(u|v)\log p_{U|V}(u|v)\, du\right\}, \tag{149}$$

where $p_{U|V}(u|v) = p_{UV}(u, v)/p_V(v)$ (when $p_V(v) > 0$) is the conditional density for U given V. Of course, we can write

$$H(U|V) = \int_{-\infty}^{\infty} p_V(v)H(U|V = v)\, dv,$$

where $H(U|V = v)$ is the term in brackets in (149). Furthermore, the Shannon information between U and V is often taken as

$$I(U; V) = H(U) - H(U|V). \tag{150}$$

In this aection we need to define $H(U|V)$ for a general random quantity V—in particular, an infinite aequence of random variables. The simpleat way to eliminate detailed mathematical technicalities is to exploit the fact that the information $I(U; V)$ for abstract random quantities U, V has been carefully defined, and many of its properties established in the literature.[5,6] We then define $H(U|V)$ in terms of $I(U; V)$ and $H(U)$ using (150). Thus, for U a real-valued random variable such that $H(U) < \infty$, and V an arbitrary random quantity, the *conditional (differential) entropy* of U given V is defined as

$$H(U|V) \triangleq H(U) - I(U; V). \tag{151}$$

Well-known properties of the Shannon information[4,5] can now be used to verify these additional facts:

$$(d)\ H(U|V) = H(U), \quad U, V \text{ independent}, \tag{152}$$

$$(e)\quad H(U\,|\,V) \geq H(U\,|\,V, W), \tag{153}$$

$$(f)\quad H(U + f(V)\,|\,V) = H(U\,|\,V), \tag{154}$$

$$(g)\quad H(U\,|\,V, f(V)) = H(U\,|\,V), \tag{155}$$

where U is any random variable with $H(U) < \infty$, V, W are any random quantities, and f is any measurable function.

Finally, let $U = \{U_n\}_{n=-\infty}^{\infty}$ be a stationary real-valued random sequence. The entropy of the stationary sequence U is defined by†

$$H(U) \triangleq H(U_n\,|\,U_{-\infty}^{n-1}) = H(U_n\,|\,U_{n+1}^{\infty}). \tag{156}$$

We will be concerned with the entropy of a certain family of random sequences defined as follows. Let $Y = \{Y_n\}$ be a real-valued stationary random sequence such that $E\,|\,Y_n\,|^s < \infty$, for some $s > 0$. We are interested in the random sequence $X = \{X_n\}$ defined by

$$X_n = \sum_{m=0}^{M} a_m Y_{n-m}, \quad -\infty < n < \infty, \tag{157}$$

where $M < \infty$ and for convenience $a_0 = 1$. Since $E\,|\,Y_n\,|^s$, $E\,|\,X_n\,|^s < \infty$, the entropies $H(Y) = H(Y_n\,|\,Y_{-\infty}^{n-1})$ and $H(X) = H(X_n\,|\,X_{-\infty}^{n-1})$ are meaningful. Our first task is to establish a relation between $H(X)$ and $H(Y)$.

The polynomial

$$Q(z) \triangleq \sum_{m=0}^{M} a_m z^m = \prod_{j=1}^{M} (1 - \alpha_j z) \tag{158}$$

is associated with the process X. The $\{\alpha_j^{-1}\}_{j=1}^{M}$ are the M, perhaps complex and/or repeated roots of $Q(z)$.

The main result of this section is a theorem which relates the entropy $H(X)$ of the stationary sequence X to the entropy $H(Y)$ of the sequence Y. After giving the proof, we show how to apply it to obtain a bound on the prediction error.

---

† The second equality of (156) follows from

$$H(U_n\,|\,U_{-\infty}^{n-1}) = \lim_{N\to\infty} H(U_n\,|\,U_{n-N}^{n-1}), \quad \text{and (see Ref. 5)}$$

$$\begin{aligned} I(U_n;\, U_{n-N}^{n-1}) &= I(U_n;\, U_{n-1}) + I(U_n;\, U_{n-2}\,|\,U_{n-1}) \\ &\quad + \cdots + I(U_n;\, U_{n-N}\,|\,U_{n-N+1}^{n-1}) \\ &= I(U_n;\, U_{n+1}) + I(U_n;\, U_{n+2}\,|\,U_{n+1}) \\ &\quad + \cdots + I(U_n;\, U_{n+N}\,|\,U_{n+1}^{n+N-1}) \\ &= I(U_n;\, U_{n+1}^{n+N}). \end{aligned}$$

Theorem 2:

(*a*)

$$H(X) \geq H(Y) + \sum_{j:|\alpha_j|>1} \log|\alpha_j|$$

$$\triangleq H(Y) + \log \Delta. \tag{159}$$

(*b*) *If Y is ergodic and $E|Y_n| < \infty$, or if $|\alpha_j| \neq 1$, $1 \leq j \leq M$, then* (159) *holds with equality.*

(*c*) *We exhibit a nonergodic Y for which* (159) *holds with strict inequality.*

Remark: A straightforward integration yields

$$\int_0^1 \log|Q(e^{i2\pi f})|\, df = \sum_{j=1}^{M} \int_0^1 \log|1 - \alpha_j e^{i2\pi f}|\, df$$

$$= \sum_{j:|\alpha_j|>1} \log|\alpha_j| = \log \Delta. \tag{160a}$$

Kanter[2] showed that when $\{Y_n\}$ are i.i.d.,

$$H(X) \geq H(Y) + \int_0^1 \log|Q(e^{i2\pi f})|\, df. \tag{160b}$$

Of course, when the $\{Y_n\}$ are i.i.d., Theorem 2(*b*) implies that (160b) holds with equality.†

Proof: Let us factor $Q(z)$ into $Q(z) = Q_1(z)Q_2(z)$, where

$$Q_1(z) = \prod_{j=1}^{M_1} (1 - \beta_j z) = \sum_{m=0}^{M_1} b_m z^m, \tag{161a}$$

$$Q_2(z) = \prod_{j=1}^{M_2} (1 - \gamma_j z) = \sum_{m=0}^{M_2} c_m z^m, \tag{161b}$$

where $|\beta_j| \geq 1$, $1 \leq j \leq M_1$, and $|\gamma_j| < 1$, $1 \leq \gamma \leq M_2$. Thus $Q_1(z)$ corresponds to the roots of $Q(z)$ inside and on the unit circle, and $Q_2(z)$ to the roots of $Q(z)$ outside the unit circle. The $\{\beta_j\}$ $\{\gamma_j\}$ may be complex, but tbe $\{b_m\}$ and $\{c_m\}$ are real. Of course,

$$\log \Delta = \sum_{j:|\alpha_j|>1} \log|\alpha_j| = \sum_{j=1}^{M_1} \log|\beta_j|$$

$$= \log \prod_{j=1}^{M_1} |\beta_j| = \log|b_{M_1}|. \tag{162}$$

† Let us remark that Shannon, in his classic paper ("A Mathematical Theory of Communication," *B.S.T.J.*, *27* (1948)), stated that (160b) always holds with equality, and gave an intuitive justification for this. We now know, however, that the equality will hold only if conditions, such as those in part (*b*), also hold.

Next define the delay operator D. Let $u = \{u_n\}_{n=-\infty}^{\infty}$ be a sequence. Then $(Du)_n = u_{n-1}$. Thus X can be written as $X = Q(D)Y = Q_1(D)Q_2(D)Y$. Let $W = Q_2(D)Y$, i.e.,

$$W_n = \sum_{m=0}^{M_2} c_m Y_{n-m}, \quad c_0 = 1. \tag{163}$$

Then

$$\begin{aligned} H(W) &= H(W_n \mid W_{-\infty}^{n-1}) \overset{(1)}{\geq} H(W_n \mid W_{-\infty}^{n-1}, Y_{-\infty}^{n-1}) \\ &= H(Y_n + \sum_{m=1}^{M_2} c_m Y_{n-m} \mid W_{-\infty}^{n-1}, Y_{-\infty}^{n-1}) \\ &\overset{(2)}{=} H(Y_n \mid W_{-\infty}^{n-1}, Y_{-\infty}^{n-1}) \overset{(3)}{=} H(Y_n \mid Y_{-\infty}^{n-1}) = H(Y). \end{aligned} \tag{164}$$

Step (1) follows from (153), step (2) from (154), and step (3) from (155), since $W_{-\infty}^{n-1}$ can be calculated from $Y_{-\infty}^{n-1}$ using (163).

We now relate $H(X)$ to $H(W)$ using the relation $X = Q_1(D)W$, i.e.,

$$X_n = \sum_{m=0}^{M_1} b_m W_{n-m}.$$

Write

$$\begin{aligned} H(X) &= H(X_n \mid X_{n+1}^{\infty}) \overset{(1)}{\geq} H(X_n \mid X_{n+1}^{\infty}, W_{n-M_1+1}^{\infty}) \\ &= H\left(\sum_{m=0}^{M_1-1} b_m W_{n-m} + b_{M_1} W_{n-M_1} \mid X_{n+1}^{\infty}, W_{n-M_1+1}^{\infty}\right) \\ &\overset{(2)}{=} H(b_{M_1} W_{n-M_1} \mid X_{n+1}^{\infty}, W_{n-M_1+1}^{\infty}) \\ &\overset{(3)}{=} H(b_{M_1} W_{n-M_1} \mid W_{n-M_1+1}^{\infty}) \\ &\overset{(4)}{=} \log \mid b_{M_1} \mid + H(W_{n-M_1} \mid W_{n-M_1+1}^{\infty}) \\ &= \log \mid b_{M_1} \mid + H(W) \overset{(5)}{=} \log \Delta + H(W). \end{aligned} \tag{165}$$

Step (1) follows from (153), step (2) from (154), step (3) from (155) and the fact that X_{n+1}^{∞} can be calculated from $W_{n-M_1+1}^{\infty}$, step (4) from (148), and step (5) from (162). Combining (164) and (165), we obtain $H(X) \geq H(Y) + \log \Delta$, which is (159), i.e., part (a) of the theorem.

We now inquire about the conditions under which (159) holds with equality. Clearly, this will happen if steps (1) in relations (164) and (165) hold with equality. From (155), this occurs if there exist measur-

able functions g_1, g_2 such that

$$Y_{-\infty}^{n-1} = g_1(W_{-\infty}^{n-1}) \quad \text{a.s.}, \tag{166a}$$

and

$$W_{n-M_1+1}^{\infty} = g_2(X_{n+1}^{\infty}) \quad \text{a.s.} \tag{166b}$$

Hence we shall show that when the conditions of part (*b*) are satisfied, then (166) will hold.

We begin by considering the following simple situation. Let $U = \{U_n\}$ be an arbitrary real-valued stationary random sequence, such that $E|U_n|^s < \infty$, for some $s > 0$. Let the random sequence $V = \{V_n\}$ be defined by

$$V_n = U_n - \xi U_{n-1}, \quad -\infty < n < \infty, \tag{167}$$

where ξ is a complex number. We can write, for $N = 1, 2, \cdots$,

$$U_n = \sum_{k=0}^{N-1} \xi^k V_{n-k} + \xi^N U_{n-N}, \quad -\infty < n < \infty. \tag{168}$$

We now show that when $|\xi| < 1$, $\xi^N U_{n-N} \to 0$, as $N \to \infty$, a.s. This will follow when we show that for any $\epsilon > 0$, for $N = 1, 2, \cdots$,

$$P(|\xi^N U_N| > \epsilon, \text{ i.o.}) = 0. \tag{169}$$

To establish (169), invoke the Borel-Cantelli lemma and write

$$\sum_{N=1}^{\infty} P(|\xi^N U_N| > \epsilon) = \sum_{N=1}^{\infty} P(|U_N| > \epsilon |\xi|^{-N})$$

$$\leq \sum_{N=1}^{\infty} \frac{E|U_N|^s}{(\epsilon|\xi|^{-N})^s} < \infty.$$

In (168), we let $N \to \infty$, and conclude that, when $|\xi| < 1$,

$$U_n = \sum_{k=0}^{\infty} \xi^k V_{n-k}, \quad \text{a.s.}, \tag{170a}$$

and that

$$U_{-\infty}^{n} = f_1(V_{-\infty}^{n}), \quad \text{a.s.}, \tag{170b}$$

where f_1 is the function defined by (170a).

Return now to the random sequences X, Y related by $X = Q_1(D)Q_2(D)Y = Q_1(D)W$, where $W = Q_2(D)Y$. Using (161b), we have

$$W = \prod_{j=1}^{M_2} (1 - \gamma_j D)Y, \quad |\gamma_j| < 1.$$

Invoking the result of (170) M_2 times, we conclude that $Y_{-\infty}^{n-1}$ is a.s. calculable from $W_{-\infty}^{n-1}$, i.e., (166a) holds.

We next investigate when (166b) will hold. Using (161a), we have

$$X = Q_1(D)W = \prod_{j=1}^{M_1} (1 - \beta_j D)W. \tag{171}$$

This prompts us to consider the process defined by (167) when $|\xi| \geq 1$. Rewriting (167) as

$$U_n = \xi^{-1}U_{n+1} - \xi^{-1}V_{n+1}, \quad -\infty < n < \infty$$

so that as in the derivation of (170) we have, for $|\xi| > 1$,

$$U_n = \sum_{k=1}^{\infty} \xi^{-k}V_{n+k}, \quad -\infty < n < \infty, \tag{172a}$$

so that

$$U_n = f_2(V_{n+1}^{\infty}), \quad -\infty < n < \infty. \tag{172b}$$

If all the $|\beta_j| > 1$, $1 \leq j \leq M_1$, then application of (172) to (171) M_1 times yields (166b). Thus (159) will hold with equality when all $|\beta_j| \neq 1$ or equivalently all $|\alpha_j| \neq 1$.

We complete the proof of part (*b*) by showing that if Y is ergodic and $E|Y_n| < \infty$, then (166b) holds. It will suffice to show that, with U a stationary ergodic sequence with $E|U_n| < \infty$ and V defined by (167) with $|\xi| = 1$, we can calculate U_n^{∞} from V_{n+1}^{∞}. From (167) we obtain, for $-\infty < n < \infty$, $1 \leq k < \infty$,

$$U_n = \sum_{j=1}^{k} \xi^{-j}V_{n+j} + \xi^{-k}U_{n+k}.$$

Thus†

$$U_n = \frac{1}{K}\sum_{k=1}^{K} U_n = \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{k} \xi^{-j}V_{n+j} + \frac{1}{K}\sum_{k=1}^{K} \xi^{-k}U_{n+k}.$$

We will show that, as $K \to \infty$,

$$\frac{1}{K}\sum_{k=1}^{K} \xi^{-k}U_{n+k} \to c, \quad \text{a.s.} \tag{173}$$

where c is a constant (in fact $c = 0$, $\xi \neq 1$ and $c = EU_n$, $\xi = 1$), which will imply that

$$U_n = \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{k} \xi^{-j}V_{n+j} + c,$$

completing the proof of part (*b*).

† This step was suggested by A. Gersho.

It remains to verify (173). Let θ be a complex random variable uniformly distributed on the unit circle, and independent of the sequence U. Then $\{\theta\xi^{-j}\}_{j=-\infty}^{\infty}$ is a stationary sequence, and therefore $\{\theta\xi^{-j}U_{n+j}\}_{j=-\infty}^{\infty}$ is also stationary. Thus, as $K \to \infty$,

$$\frac{1}{K}\sum_{j=1}^{K} \theta\xi^{-j}U_{n+j} \to \eta,$$

a random variable with probability 1, and therefore

$$\frac{1}{K}\sum_{j=1}^{K} \xi^{-j}U_{n+j} \to \eta\theta^{-1} \triangleq c, \quad \text{a.s.} \tag{174}$$

Since the left member of (174) depends only on the tail σ-field of the $\{U_n\}$, and not on θ, we conclude that c is a constant a.s. In fact, since the expectation

$$E\frac{1}{K}\sum_{j=1}^{K} \xi^{-j}U_{n+j} = \begin{cases} 0, & \xi \neq 1, \\ EU_n, & \xi = 1, \end{cases}$$

we know the value of c.

Our final task, part (c) of Theorem 2, is to exhibit a situation in which $H(X) > H(Y) + \log \Delta$. Let

$$X_n = Y_n - Y_{n-1},$$

and let

$$Y_n = \eta_n + \nu$$

where $\{\eta_n\}_{n=-\infty}^{\infty}$ are i.i.d. standard Gaussian variates, and ν is a random variable defined as follows. Let

$$\epsilon_{2j} = \begin{cases} 1 & \eta_j > 0, \\ 0 & \eta_j \leq 0, \end{cases} \qquad j = 0, 1, 2, \cdots$$

and let

$$\epsilon_{2j-1} = \begin{cases} 1 & \eta_{-j} > 0, \\ 0 & \eta_{-j} \leq 0, \end{cases} \qquad j = 1, 2, \cdots.$$

Then the binary expansion of ν is $0.\epsilon_0\epsilon_1\epsilon_2\epsilon_3\cdots$. Thus knowledge of ν is equivalent to knowledge of the sign of η_n, $-\infty < n < \infty$. Note that the sequence Y is stationary but not ergodic. Also $\log \Delta = 0$, so that (159) is $H(X) \geq H(Y)$. Now

$$X_n = \eta_n - \eta_{n-1},$$

and $\{\eta_n\}$ is an ergodic sequence. Thus by part (b) $H(X) = H(\eta) = H(\eta_n) = \frac{1}{2}\log 2\pi e$, the last equality following by direct integration. We now consider $H(Y)$. Observe that

$$\frac{1}{N}\sum_{k=1}^{N} Y_{n+k} = \left(\frac{1}{N}\sum_{k=1}^{N} \eta_{n+k}\right) + \nu \to \nu, \quad \text{a.s.}$$

as $N \to \infty$. Thus ν and therefore $\epsilon = \{\epsilon_j\}_{j=0}^{\infty}$, are calculable from Y_{n+1}^{∞}. It follows that

$$H(Y) = H(Y_n \mid Y_{n+1}^{\infty}) = H(Y_n \mid Y_{n+1}^{\infty}, \epsilon, \nu) = H(\eta_n \mid Y_{n+1}^{\infty}, \epsilon, \nu)$$

$$\leq H(\eta_n \mid \epsilon) = \frac{1}{2} \log \frac{\pi e}{2} < \frac{1}{2} \log 2\pi e = H(X).$$

This establishes part (c) and completes the proof of Theorem 2. Let us remark that it is possible, using the Cantor diagonalization method, to imbed a complete specification of the $\{\eta_n\}$ in ν, thus making $H(Y) = -\infty$, without changing $H(X)$.

Our next task is the application of the theorem to our estimation problem. Let $X = \{X_n\}$ be any stationary process such that $H(X_n) < \infty$. Let $\hat{X}_n = f(X_{-\infty}^{n-1})$ be an estimator of X_n, and let the figure of merit be $E \mid X_n - \hat{X}_n \mid^r$, where $r > 0$. It follows that

$$I(X_n; X_{-\infty}^{n-1}) \overset{(1)}{\geq} I(X_n; \hat{X}_n) = H(X_n) - H(X_n \mid \hat{X}_n)$$

$$\overset{(2)}{\geq} H(X_n) - \frac{1}{r} \log \frac{2^r e \Gamma(1/r) E \mid X_n - \hat{X}_n \mid^r}{r^{r-1}}, \quad (175)$$

where step (1) follows from the data-processing theorem (which states that processing $X_{-\infty}^{n-1}$ to form $\hat{X}_n$ decreases information),[6,4] and step (2) from (146) and the concavity of the logarithm. Since $I(X_n; X_{-\infty}^{n-1}) = H(X_n) - H(X)$, (175) yields

$$E \mid X_n - \hat{X}_n \mid^r \geq \frac{r^{(r-1)} e^{rH(X)}}{2^r e \Gamma^r(1/r)}. \quad (176a)$$

In the special case $r = 2$, (176) becomes

$$E \mid X_n - \hat{X}_n \mid^2 \geq \frac{e^{2H(X)}}{2\pi e}. \quad (176b)$$

Inequalities (176) hold for any stationary process X. When X is given in terms of another process Y by (157), we can use part (a) of Theorem 2 to continue these inequalities, i.e.,

$$E \mid X_n - \hat{X}_n \mid^r \geq \frac{r^{(r-1)} \Delta^r e^{rH(Y)}}{2^r e \Gamma^r(1/r)}, \quad (177a)$$

and, for $r = 2$,

$$E \mid X_n - \hat{X}_n \mid^2 \geq \frac{\Delta^2}{2\pi e} e^{2H(Y)}, \quad (177b)$$

where Δ is given in (160a).

Inequality (177b) has an interesting interpretation. Assume that the spectral density of Y,

$$S_Y(f) = \sum_{n=-\infty}^{\infty} \rho_Y(n) e^{i2\pi nf}, \tag{178}$$

where $\rho_Y(n) = EY_m Y_{m+n}$, $-\infty < n < \infty$, exists. Then the spectral density of X is

$$S_X(f) = S_Y(f) |Q(e^{i2\pi f})|^2.$$

Using (160a) and the well-known formula for $\epsilon^{*2}_{\infty \text{lin}}$[1]

$$2 \log \Delta = \int_0^1 \log S_X(f)\, df - \int_0^1 \log S_Y(f)\, df$$

$$= \log \epsilon^{*2}_{\infty \text{lin}} - \int_0^1 \log S_Y(f)\, df,$$

so that

$$\Delta^2 = \frac{\epsilon^{*2}_{\infty \text{lin}}}{\bar{S}_Y}. \tag{179}$$

Here $\epsilon^{*2}_{\infty \text{lin}}$ is, as in Section I, the best linear mean-squared prediction error, and $\bar{S}_Y$ is the geometric mean of the spectral density of Y. Substituting (179) into (177b), we have

$$E |X_n - \hat{X}_n|^2 \geq \epsilon^{*2}_{\infty \text{lin}} [e^{2H(Y)}/(2\pi e \bar{S}_Y)]. \tag{180}$$

Now it is not hard to show that when U is a Gaussian process with spectral density S_U, $H(U) = \frac{1}{2} \log (2\pi e \bar{S}_U)$. Thus $e^{2H(Y)}/2\pi e$ is the geometric mean of the spectral density of a Gaussian process with the same entropy as Y. This is called the "entropy power" of Y. Thus the quantity in brackets in (180) is the ratio of the entropy power of Y to $\bar{S}_Y$, and is unity when Y is Gaussian. Kanter[2] obtained (180) when the $\{Y_n\}$ are i.i.d. We have proved it for any stationary Y_n with $E |Y_n|^s < \infty$ for some $s > 0$.

VII. ACKNOWLEDGMENT

We gratefully acknowledge the help of P. Diaconis, B. F. Logan, and C. L. Mallows in obtaining the results in Appendix B, and A. Gersho for his contribution to Theorem 2. We particularly thank M. Kanter for arousing our interest in this problem by a talk given at Bell Laboratories on his earlier cited work.

APPENDIX A

Let

$$X_n = Y_n - aY_{n-1}, \quad EY_n = 0, \quad EY_n^2 = 1 \tag{181}$$
$$n = 0, \pm 1, \pm 2, \cdots$$

with the Y's i.i.d. variates. For $a > 0$, let

$$f_k^*(x_1, x_2, \cdots, x_k; a)$$
$$\equiv E(X_n \mid X_{n-1} = x_1, X_{n-2} = x_2, \cdots, X_{n-k} = x_k), \tag{182}$$
$$\epsilon_k^{*2}(a) \equiv E[X_n - E(X_n \mid X_{n-1}, X_{n-2}, \cdots, X_{n-k})]^2. \tag{183}$$

Theorem 3: If the Y's in (181) are symmetric, i.e., if $P_Y(y) = P_Y(-y)$, and if for $a \geq 0$ (182) and (183) hold, then for $a \leq 0$

$$E(X_n \mid X_{n-1} = x_1, \cdots, X_{n-k} = x_k)$$
$$= f_k^*(-x_1, x_2, -x_3, \cdots, (-1)^k x_k; \mid a \mid),$$
$$E[X_n - E(X_n \mid X_{n-1}, \cdots, X_{n-k})]^2 = \epsilon_k^{*2}(\mid a \mid).$$

Proof: In (181) set

$$X_n' = (-1)^{n+a} X_n, \qquad Y_n' = (-1)^{n+a} Y_n \qquad n = 0, \pm 1, \cdots \tag{184}$$

for some integer a. In terms of these new variables, (181) becomes

$$X_n' = Y_n' + aY_{n-1}', \qquad EY_n' = 0, \qquad EY_n'^2 = 1.$$

Furthermore $P_Y'(y) = P_Y(y)$. Thus if $a = -b < 0$ the X_n' have the same distribution as the X_n do with $a = b$ in (181). Thus if $a \leq 0$

$$E(X_n' \mid X_{n-1}' = x_1, \cdots, X_{n-k}' = x_k) = f_k^*(x_1, x_2, \cdots, x_k; \mid a \mid)$$

Now in (184) let $a = -n$, so that $X_n' = X_n$, $X_{n-1}' = -X_{n-1}$, $X_{n-2}' = X_{n-2}$, etc. But then $E(X_n' \mid X_{n-1}' = x_1, \cdots, X_{n-k}' = x_k) = E(X_n \mid X_{n-1} = -x_1, X_{n-2} = x_2, \cdots, X_{n-k} = (-1)^k x_k)$ and the theorem readily follows.

APPENDIX B

Let Y_0 and Y_1 be i.i.d. with $EY_i^2 < \infty$. We shall show that

$$E \operatorname{var}(Y_1 \mid Y_1 - Y_0) \leq E \operatorname{var}(Y_1 \mid Y_1 + Y_0). \tag{185}$$

This says that, in estimating Y_1 in the mean-square sense, the average error is less if the difference $Y_1 - Y_0$ is known than if the sum is known. If Y_0 and Y_1 are replaced by Y_{n-1} and Y_n where $X_n = Y_n - aY_{n-1}$ as in Section I, then (185) states that $\epsilon_1^{*2}(a = -1) \leq \epsilon_1^{*2}(a = +1)$, as was asserted in Section I. We prove (185) below and the assertion that equality holds in (185) only if Y_i are symmetric *provided* that Y_0 has a characteristic function $\phi(t) = Ee^{iY_0 t}$ which is nowhere zero. However, in general there are nonsymmetric Y_i for which equality holds in (185).

The following proof of (185) was suggested by C. Mallows: Assume without loss of generality that $EY_i = 0, EY_i^2 = 1$. Observe first that the right side of (185) is ½ because

$$E(Y_1 \mid Y_1 + Y_0) = E(Y_0 \mid Y_1 + Y_0) = E\left[\frac{Y_1 + Y_0}{2} \mid Y_1 + Y_0\right] = \frac{Y_1 + Y_0}{2}$$

and

$$E\left(Y_1 - \frac{Y_1 + Y_0}{2}\right)^2 = \frac{1}{2}.$$

Set $f(z) = E(Y_1 + Y_0 \mid Y_1 - Y_0 = z)$ and note that

$$Ef(Y_1 - Y_0) = 0,$$

$$E(Y_1 - Y_0)f(Y_1 - Y_0) = E(Y_1 - Y_0)(Y_1 + Y_0) = 0. \tag{186}$$

Now

$$E[Y_1 \mid Y_1 - Y_0] = E\left[\frac{Y_1 - Y_0}{2} + \frac{Y_1 + Y_0}{2} \mid Y_1 - Y_0\right]$$

$$= \frac{Y_1 - Y_0}{2} + \frac{1}{2} f(Y_1 - Y_0). \tag{187}$$

Note that

$$E \operatorname{var}[Y_1 \mid Y_1 - Y_0] = EE[(Y_1 - E[Y_1 \mid Y_1 - Y_0])^2 \mid Y_1 - Y_0]$$

$$= E(Y_1 - E(Y_1 \mid Y_1 - Y_0)^2 = 1 - E(E[Y_1 \mid Y_1 - Y_0])^2. \tag{188}$$

But using (186) and (187),

$$E(E[Y_1 \mid Y_1 - Y_0])^2 = \operatorname{var}(E[Y_1 \mid Y_1 - Y_0])$$

$$= E\left(\frac{Y_1 - Y_0}{2}\right)^2 + E\left(\frac{1}{2} f(Y_1 - Y_0)\right)^2 \geq \frac{1}{2} \tag{189}$$

and (185) is proved.

To prove the assertions about equality, let $\phi(t) = Ee^{iY_t}$ and note that

$$\phi(u - v)\phi(u + v) = EE[e^{i(Y_0 + Y_1)u} \mid Y_0 - Y_1]e^{i(Y_0 - Y_1)v}. \tag{190}$$

Now equality holds in (189) if and only if $f(z) \equiv 0$ which is the same as requiring that the derivative of the left side of (190) at $u = 0$ vanish for all v. That is,

$$\phi'(v)\phi(-v) + \phi'(-v)\phi(v) \equiv 0, \quad -\infty < v < \infty. \tag{191}$$

If $\phi \neq 0$, this says that $(\log \phi(v))' = \phi'(v)/\phi(v)$ is odd and so $\log \phi(v)$ and $\phi(v)$ are even, i.e., Y is symmetric. Note that a characteristic function is real if and only if the corresponding random variable is symmetric.

Our aim now is to show that (191) can hold without ϕ being real. We shall construct a $\phi_0(t)$ that satisfies (191) and is positive definite, real for $|t| < 1$, and pure imaginary for $|t| \geq 1$. We write

$$\phi_0(t) = \phi_1(t) + \epsilon\phi_2(t) \tag{192}$$

and denote Fourier transforms by upper-case letters so that

$$\phi_j(t) = \int_{-\infty}^{\infty} \Phi_j(x) e^{ixt}\, dx, \qquad \Phi_j(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_j(t) e^{-ixt}\, dt \tag{193}$$

$$j = 0, 1, 2.$$

Now choose $\phi_2(t) \not\equiv 0$ to be purely imaginary, to be six times continuously differentiable and of compact support, and to satisfy

$$\phi_2(t) \not\equiv 0, \qquad |t| \leq 1, \qquad \phi_2(t) = -\phi_2(-t). \tag{194}$$

From (193) one readily finds that

$$\Phi_2(x) = \Phi_2^*(x) \tag{195}$$

and that there exists a positive constant c_1 such that

$$|\Phi_2(x)| \leq \frac{c_1}{(1 + x^2)^3}. \tag{196}$$

We next choose

$$\Phi_1(x) = c_2[H^2(x - 1) + H^2(x + 1)] \tag{197}$$

where

$$H(x) = \frac{1}{4\pi}\left[\frac{\sin(x/4)}{(x/4)}\right]^2. \tag{198}$$

Here $c_2 > 0$ is a constant chosen to make

$$\phi_0(0) = \phi_1(0) = \int_{-\infty}^{\infty} \Phi_1(x)\, dx = 1. \tag{199}$$

Note that $\Phi_1(x) > 0$ for all x and that, for all x, $x^4\Phi_1(x) \geq c_3 > 0$. Since from (195) $\Phi_2(x)$ is real and from (196) is $0(1/x^6)$, it is possible to choose ϵ sufficiently small so that

$$\Phi_0(x) = \Phi_1(x) + \epsilon\Phi_2(x) > 0$$

for all x. ϕ_0 is thus the transform of a real positive function and satisfies the normalization (199). It is therefore a characteristic function.

We next note that

$$\phi_1(t) = 0, \qquad |t| > 1. \tag{200}$$

This follows immediately from the fact that

$$h(t) = \int_{-\infty}^{\infty} H(x)e^{ixt}\,dx = \begin{cases} 1 - 2|t|, & |t| \le \tfrac{1}{2} \\ 0, & |t| > \tfrac{1}{2} \end{cases}$$

has support $|t| \le \frac{1}{2}$ so that the transforms of $H^2(x-1)$ and of $H^2(x+1)$ both have support $|t| \le 1$. Equation (197) then implies (200). The form of (197) also shows that

$$\phi_1(t) = \phi_1(-t) \tag{201}$$

and that $\phi_1(t)$ is twice differentiable.

We have now constructed a twice differentiable characteristic function $\phi_0(t) = \phi_1(t) + \epsilon\phi_2(t)$ where real even ϕ_1 vanishes for $|t| > 1$ and pure imaginary odd ϕ_2 vanishes for $|t| \le 1$. For all t, then, where $\phi_0(t) \neq 0$, the quantity $\phi_0'(t)/\phi_0(t)$ is real and odd. Thus $\phi_0'(t)/\phi_0(t) = (\phi_0'(t)/\phi_0(t))^* = -\phi_0'(-t)/\phi_0(-t)$. That is, the characteristic function ϕ_0 satisfies (191) and is not real (for $t > 1$). QED.

Note added in print: We have recently learned of the work of M. Rosenblatt,[7] which overlaps the present paper slightly.

REFERENCES

1. A. M. Yaglom, *An Introduction to the Theory of Stationary Random Functions*, Englewood Cliffs, N.J.: Prentice Hall, 1962, pp. 97–125.
   U. Grenander and M. Rosenblatt, *Statistical Analysis of Stationary Time Series*, New York: John Wiley, 1957, pp. 64–82.
2. M. Kanter, "Lower Bounds for Nonlinear Prediction Error in Moving-Average Processes," Ann. Probability, *7* (February 1979), pp. 128–138.
3. R. B. Ash, *Information Theory*, New York: Interscience, 1965, Ch. 8.
4. A. D. Wyner and J. Ziv, "On Communication of Analog Data from a Bounded Source Space," B.S.T.J., *48* No. 10 (December 1969), pp. 3139–3172, Theorem 1.
5. M. Pinsker, *Information and Information Stability of Random Variables and Processes*, San Francisco: Holden-Day, 1964.
6. A. D. Wyner, "A Definition of Conditional Mutual Information for Arbitrary Ensembles," Information and Control, *38* (July 1978), pp. 51–59.
7. M. Rosenblatt, *Linearity and Nonlinearity in Time Series*, Preprint.